AGU PUBLICATIONS

Geophysical Research Letters

RESEARCH LETTER
10.1002/2017GL073798

Key Points:
- Radiosonde observations show good agreement with a single-model 100 member ensemble of simulations for the majority of observational record
- Largest trend differences for amplification of tropospheric over surface warming aloft for 30-year trends occur in second half of the record
- Differences in observed and simulated tropical tropospheric temperature trends arise from observational uncertainty and internal variability

Supporting Information:
- Supporting Information S1

Correspondence to:
L. Suárez-Gutiérrez,
laura.suarez@mpimet.mpg.de

Citation:
Suárez-Gutiérrez, L., C. Li, P. W. Thorne, and J. Marotzke (2017), Internal variability in simulated and observed tropical tropospheric temperature trends, *Geophys. Res. Lett.*, *44*, 5709–5719, doi:10.1002/2017GL073798.

Received 4 NOV 2016
Accepted 19 MAY 2017
Accepted article online 22 MAY 2017
Published online 11 JUN 2017

Internal variability in simulated and observed tropical tropospheric temperature trends

Laura Suárez-Gutiérrez[1], Chao Li[1], Peter W. Thorne[2], and Jochem Marotzke[1]

[1]Max-Planck-Institut für Meteorologie, Hamburg, Germany, [2]Department of Geography, Maynooth University, Maynooth, Ireland

Abstract We explore the extent to which internal variability can reconcile discrepancies between observed and simulated warming in the upper tropical troposphere. We compare all extant radiosonde-based estimates for the period 1958–2014 to simulations from the Coupled Model Intercomparison Project phase 5 multimodel ensemble and the 100 realization Max Planck Institute large ensemble. We consider annual mean temperatures and all available 30-and 15-year trends. Most observed trends fall within the ensemble spread for most of the record, and trends calculated over 15-year periods show better agreement than 30-year trends, with generally larger discrepancies for the older observational products. The simulated amplification of surface warming aloft in the troposphere is consistent with observations, and the linear correlation between surface and simultaneous tropospheric warming trends decreases with trend length. We conclude that trend differences between observations and simulations of tropical tropospheric temperatures are dominated by observational uncertainty and chaotic internal variability rather than by systematic errors in model performance.

1. Introduction

General Circulation Models (GCMs) robustly simulate the vertical amplification of tropical (20°N–20°S) tropospheric warming in response to increasing concentrations of well-mixed greenhouse gases, with a warming maximum at the upper troposphere level around 300 hPa. This characteristic has important implications for climate sensitivity, due to its relation with lapse-rate and associated water vapor and cloud feedbacks, as well as with atmospheric circulation changes [*Manabe and Stouffer*, 1980; *Colman*, 2001; *Flato et al.*, 2013]. During recent decades, it has been intensively discussed whether the simulated response adequately captures the real-world behavior described by observations [*Intergovernmental Panel on Climate Change*, 2013]. Discrepancies between observed and simulated tropical tropospheric temperature structures caused by incorrect model performance would compromise the reliability and accuracy of model-based projections for future climate change. The vertical temperature structure of the troposphere, as any climate parameter, emerges from a superposition of natural internal variability and response to external forcings. The observational record represents only one outcome of many possible resulting from this combination. We use two large ensembles of different realizations—with each realization simulating also one outcome of this superposition—to assess to what extent internal variability can reconcile apparent discrepancies.

Most model simulations show larger warming trends in the tropical troposphere than observational data sets over recent years [*McKitrick et al.*, 2010; *Fu et al.*, 2011; *Po-Chedley and Fu*, 2012]. Some studies argue that models significantly overestimate the tropospheric warming response and are therefore seriously flawed [*Douglass et al.*, 2007; *McKitrick et al.*, 2010]. Other studies, however, suggest that remaining discrepancies are not statistically significant once observational uncertainties and natural internal variability are taken into account [*Sherwood et al.*, 2005; *Thorne et al.*, 2007; *Santer et al.*, 2008; *Thorne et al.*, 2011a, 2011b; *Mitchell et al.*, 2013; *Sherwood and Nishant*, 2015; *Po-Chedley et al.*, 2015]. Yet other studies cannot unambiguously conclude whether discrepancies are caused by incorrect model behavior or observational uncertainties [*Fu et al.*, 2011; *Po-Chedley and Fu*, 2012; *Seidel et al.*, 2012; *Santer et al.*, 2016].

However, most of the existing studies focus on periods starting after year 1979—the beginning of the *satellite era*—and analyze just one trend estimate [e.g., *Santer et al.*, 2005; *Douglass et al.*, 2007; *Santer et al.*, 2008; *Fu et al.*, 2011; *Po-Chedley and Fu*, 2012]. We argue that to put potential trend discrepancies into appropriate context, all overlapping trends over the longest available record of tropical tropospheric temperatures must

be analyzed [*Thorne et al.*, 2007; *Risbey et al.*, 2014; *Marotzke and Forster*, 2015]. Radiosonde-based data sets offer the longest available record and therefore are used in our study. We further evaluate the robustness of our results by comparing trend estimates under several trend calculation methods and trend lengths.

On climate time scales, tropical tropospheric temperatures are a response to surface changes amplified aloft due to the strong convective activity. Inconsistencies in model-simulated surface temperatures are expected to lead to inconsistent tropospheric temperatures resulting from this amplification [*Santer et al.*, 2005; *Mitchell et al.*, 2013]. We analyze trends of the difference between upper tropical troposphere temperature anomalies and surface temperature anomalies (TTT-minus-ST), as well as the amplification of surface warming aloft by considering upper troposphere temperature trends against simultaneous surface temperature trends (TTT-versus-ST).

We evaluate the ability of two state-of-the-art model ensembles, the Phase 5 of the Coupled Model Intercomparison Project (CMIP5) multimodel ensemble and the recently developed Max Planck Institute Earth System Model (MPI-ESM) 100 realization large ensemble, to simulate the radiosonde-observed interannual variability and trend behavior in tropical upper troposphere temperature anomalies over the period 1958–2014. To our knowledge, this is the first time that a multimember large ensemble experiment from an Earth system model has been used to assess the contribution of internal variability to CMIP5 model-simulated tropospheric warming and to examine possible discrepancies between simulations and observations in tropical tropospheric temperature anomalies.

2. Data

For observational tropospheric temperature estimates, satellite- and radiosonde-based measurements are available. We choose to use radiosonde observations, since they offer longer observational records and higher vertical resolution than satellite products, which is particularly relevant to investigating the warming maximum in the upper troposphere [*Fu et al.*, 2004; *Haimberger et al.*, 2012; *Seidel et al.*, 2012; *Sherwood and Nishant*, 2015]. Care must be taken, however, because uncorrected radiosonde data suffer from biases toward insufficient tropospheric warming, and these biases are largest in the tropics [*Lanzante et al.*, 2003; *Sherwood et al.*, 2005, 2008; *Titchner et al.*, 2009; *Thorne et al.*, 2011b]. These biases include step-like mechanisms that do not cancel over time, whereas the amplitude of internal variability diminishes with time scale. Therefore, any residual data issues will become more apparent the longer the time scale [*Santer et al.*, 2005; *Thorne et al.*, 2007; *Sherwood and Nishant*, 2015]. Although homogenization techniques remove many of these biases, some biases may remain [*Santer et al.*, 2005; *Thorne et al.*, 2007, 2011a; *Sherwood and Nishant*, 2015].

We use the most recent versions of the following six radiosonde-based compilations of tropospheric temperatures: RATPAC-A [*Free et al.*, 2005], HadAT2 [*Thorne et al.*, 2005], RAOBCORE [*Haimberger*, 2007; *Haimberger et al.*, 2012], RICH-τ [*Haimberger*, 2007; *Haimberger et al.*, 2012], RICH-obs [*Haimberger*, 2007; *Haimberger et al.*, 2012], and IUKv2 [*Sherwood and Nishant*, 2015]. The data set HadCRUT4.5 [*Morice et al.*, 2012; *Osborn and Jones*, 2014] is used as observational estimate for surface temperatures. Further details are given in the supporting information (SI) Table S1.

For simulations, we use historical and RCP4.5 runs from 44 different GCMs from the CMIP5 archive [*Taylor et al.*, 2012; *Bi et al.*, 2013; *Xin et al.*, 2013; *Ji et al.*, 2014; *von Salzen et al.*, 2013; *Meehl et al.*, 2012; *Hurrell et al.*, 2012; *Scoccimarro et al.*, 2011; *Voldoire et al.*, 2013; *Rotstayn et al.*, 2010; *Hazeleger et al.*, 2010; *Li et al.*, 2013; *Delworth et al.*, 2006; *Donner et al.*, 2011; *Schmidt et al.*, 2014; *Smith et al.*, 2010; *Collins et al.*, 2011; *Jones et al.*, 2011; *Volodin et al.*, 2010; *Dufresne et al.*, 2013; *Hourdin et al.*, 2013; *Sakamoto et al.*, 2012; *Watanabe et al.*, 2010, 2011; *Giorgetta et al.*, 2013; *Yukimoto et al.*, 2012; *Bentsen et al.*, 2013], and from the 100 realization single-model large ensemble of the MPI-ESM [*Giorgetta et al.*, 2013]. The large ensemble uses the model version MPI-ESM1.1 in low resolution (LR) configuration, with resolution T63 and 47 vertical levels in the atmosphere and 1.5° resolution and 40 vertical levels in the ocean. MPI-ESM-LR has an equilibrium climate sensitivity (ECS) of 3.6°C and transient climate response of 2°C, while the CMIP5 model mean, with its 90% uncertainty interval, is 3.2 ± 1.3°C for ECS and 1.8 ± 0.6°C for transient climate response. Information about the different CMIP5 models included in this study is listed in the SI Table S2. Most CMIP5 models include several realizations, based on the same model physics, parametrizations, and forcings. The realizations differ only in their initial conditions, taken from different points of the models preindustrial control run [*Taylor et al.*, 2012]. To better analyze the simulated internal variability in the CMIP5 multimodel ensemble, we include all individual realizations, as opposed to model or ensemble means. RCP4.5 extensions are only available for some of the corresponding

historical realizations, resulting in a reduction in the CMIP5 ensemble member number after year 2005 (SI Table S3).

We define tropical upper troposphere temperatures (TTTs) as the annual mean temperature of the 300 hPa level of the troposphere averaged over the 20°N–20°S latitude range. We choose the 300 hPa level because model simulations and most radiosonde compilations present 30-year trend warming maxima at this level; however, our results do not differ substantially for other middle-to-upper tropospheric levels. All TTTs shown are anomalies with respect to the climatology period from 1961 to 1990. When comparing the upper troposphere level to the surface, all simulated and observed surface data are subsampled to grid boxes where radiosonde data are available.

3. Results and Discussion

3.1. Internal Variability in Tropical Tropospheric Temperatures

We perform a side-by-side comparison of the variability of tropically averaged upper troposphere annual mean temperature anomalies for the period 1958–2014 simulated by the individual realizations from both ensembles against radiosonde-observed estimates. Observational estimates occur generally across the whole ensemble width and within the ensemble spread for the majority of the record, for both the CMIP5 (Figure 1a) and the MPI-ESM ensembles (Figure 1b). However, the different radiosonde series present a broad range of estimates for tropical tropospheric temperatures. This broad range indicates remaining observational uncertainties in tropical tropospheric temperatures. For TTTs over this period, we find that the spread of the 100 member MPI-ESM large ensemble is not readily distinguishable from the CMIP5 multimodel ensemble spread (Figure 1).

Observational estimates starting from 1999 occur toward the lower half of both simulated ensembles, coinciding with the observed hiatus in global surface warming [*Knight et al.*, 2010; *Wang et al.*, 2010; *Liebmann et al.*, 2010; *Flato et al.*, 2013], although observed TTTs continue to rise during this period. The year 2009 stands out in that it is the only estimate occurring outside the CMIP5 ensemble for all observational estimates except for the newest-developed IUKv2 . For the MPI-ESM large ensemble, observational TTT estimates always lie within the ensemble spread. MPI-ESM simulates TTT estimates on average 0.3°C lower than the CMIP5 ensemble for the period 2000–2014 and is therefore in slightly better agreement with observations than the CMIP5 multimodel ensemble. For this period, there is a 0.7°C TTT increase for the CMIP5 ensemble mean, while the MPI-ESM ensemble mean shows a 0.4°C increase. This difference between the CMIP5 and the MPI-ESM simulated warming may suggest that some CMIP5 models could be missing a cooling forcing contribution during this period [*Solomon et al.*, 2011; *Fyfe et al.*, 2013; *Santer et al.*, 2014], although the reduction in the CMIP5 ensemble size after year 2005 and the intermodel differences may also account for some of the difference.

The cooling following two out of three volcanic eruptions occurring during the observational record (Mount Agung, 1963, and Mount Pinatubo, 1991) is generally well represented by both ensembles at the 300 hPa level, with most models showing a slight tendency toward overestimating the cooling response and the following warming recovery. The El Chichón eruption, in 1983, presents the exception to this behavior, with observed temperatures actually rising after the volcanic eruption due to a strong El Niño event coincidentally occurring the same year as the maximum radiative cooling from the volcano, masking the volcanic cooling response [*Santer et al.*, 2001]. Both ensembles present similar responses to volcanic eruptions, with some realizations from CMIP5 showing warming responses directly following volcanic-induced cooling that are too strong, a behavior not exhibited by the MPI-ESM ensemble (Figures 1a and 1b).

The rank that the observational estimates would have as a member of each ensemble of simulations [*Hamill*, 2001; *Annan and Hargreaves*, 2010] shows a slight tendency for observational estimates to fall with a higher frequency in the vicinity of the lower simulations for the CMIP5 (Figure 1c). This behavior is not present in the MPI-ESM large ensemble (Figure 1d) and is most likely caused by estimates in the period of 1999–2014. Other than that both rank histograms show rather flat and continuous patterns, indicating that both the CMIP5 multimodel ensemble and the MPI-ESM large ensemble offer realistic representations of observed large-scale variability in TTT.

This high similarity between the 100 member MPI-ESM large ensemble—which includes just one set of external forcing contributions and forcing response—and the CMIP5 multimodel ensemble spread—which

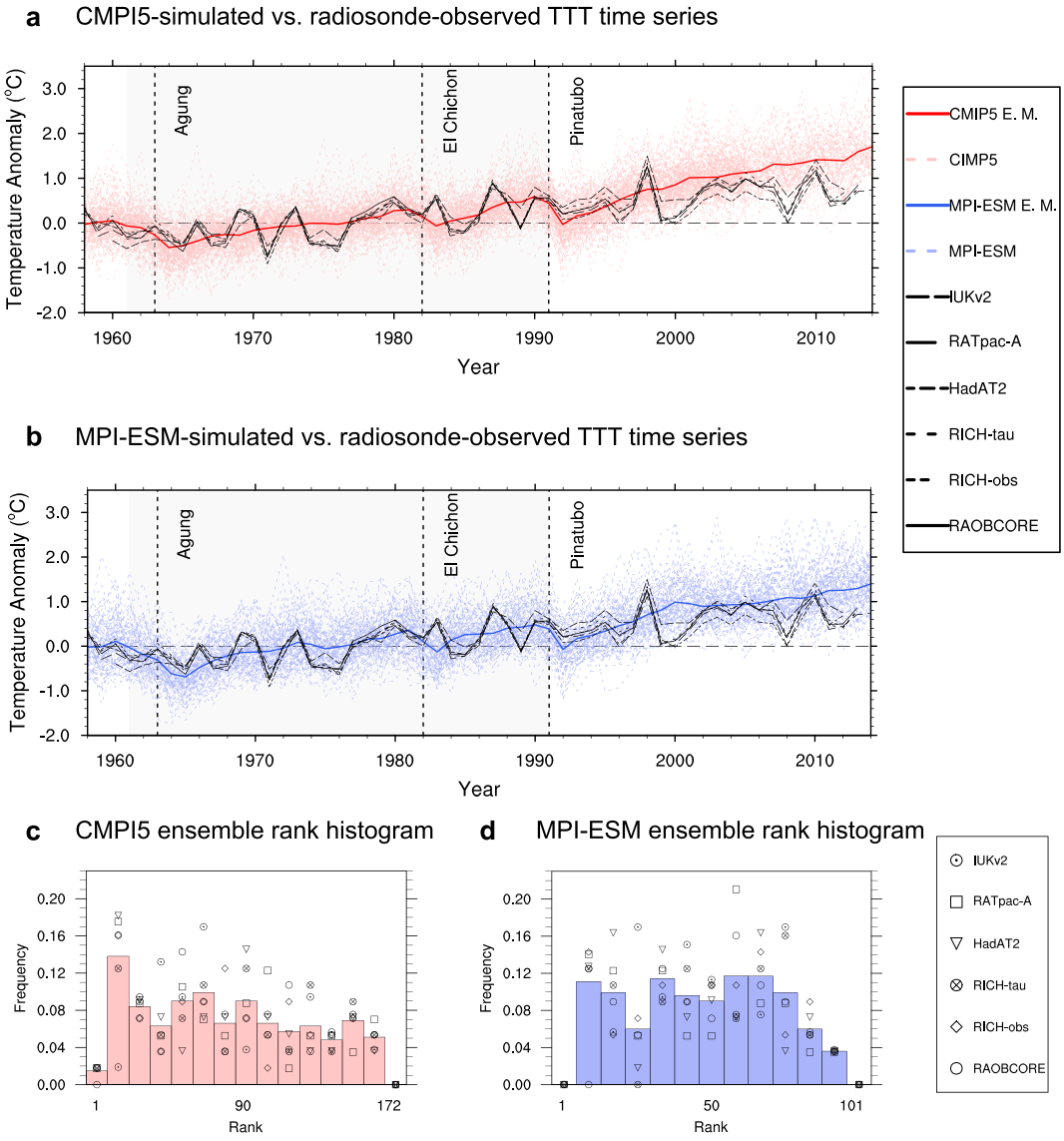

Figure 1. Simulated and radiosonde-observed TTT. (a) Radiosonde and CMIP5 time series of annual TTT at 300 hPa, for the period 1958–2014. Anomalies are calculated with respect to the period 1961–1990, shaded in gray. The tropics are defined by the 20°N–20°S range. The figure shows historical and RCP4.5 scenario runs from 44 CMIP5 models (dashed red lines), the multimodel ensemble mean (thick red line), and six radiosonde estimates (black lines). Major volcanic eruptions are marked by vertical black dashed lines. (b) Radiosonde and MPI-ESM time series of annual TTT, as in Figure 1a. The figure shows 100 historical and RCP4.5 runs from MPI-ESM (dashed blue lines), the ensemble mean (thick blue line), and radiosonde estimates (black lines). (c) Rank histogram for each radiosonde estimate as a member of the CMIP5 ensemble (black markers) and cumulative frequency for all radiosonde estimates (red bars). The frequencies are normalized to unity. (d) Rank histogram for each radiosonde estimate as a member of the MPI-ESM large ensemble (black markers) and cumulative frequency for all radiosondes (blue bars), as in Figure 1c.

includes different treatments of external forcing and different forcing responses for different GCMs — indicates that the spread in CMIP5 model-simulated TTT arises most likely mainly from simulated internal variability, with some contribution from differences in model radiative forcing and forcing response diversity. The good agreement with radiosonde TTT estimates also indicates that the CMIP5 ensemble spread offers an adequate representation of real-world variability in tropical upper tropospheric temperatures on annual time scales, as does the MPI-ESM large ensemble (Figure 1). Due to this similarity in behavior, for simplicity we focus on the MPI-ESM analysis for the rest of this paper (selected results for CMIP-5 ensemble are available in the SI Figure S1).

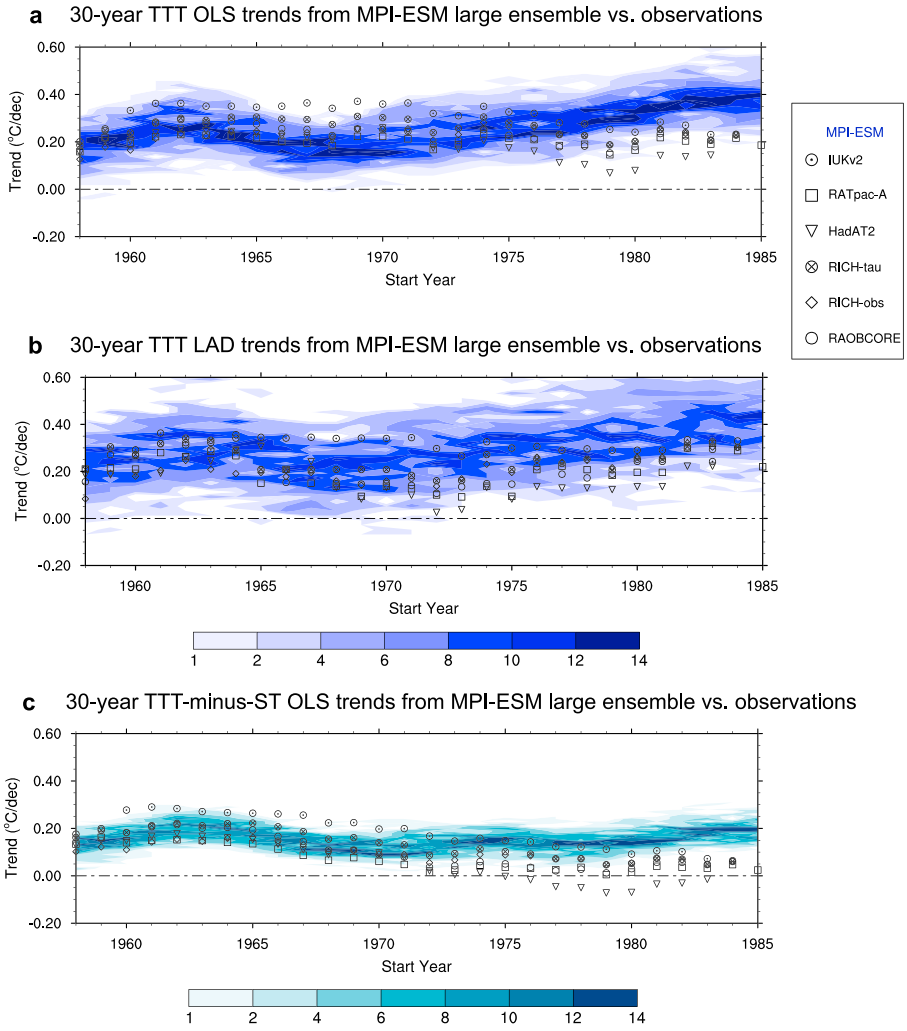

Figure 2. Simulated and observed 30-year trends. (a) Joint relative frequency distribution of MPI-ESM-simulated TTT *OLS* trend estimates compared to radiosonde estimates (black markers). Start years are separated by calendar year; bin size is 0.02°C per decade; and the color bar represents the ensemble member density, in percentage. (b) Joint relative frequency distribution of MPI-ESM-simulated TTT *LAD* trend estimates compared to radiosonde estimates, as in Figure 2a. Start years are separated by calendar year; bin size is 0.02°C per decade. (c) Joint relative frequency distribution of MPI-ESM-simulated TTT-minus-ST *OLS* trend estimates compared to radiosonde observational estimates. Bin size is 0.01°C per decade. For all radiosonde observations, the difference is shown against surface temperature values from the HadCRUT4.5 data set.

3.2. Tropical Tropospheric Temperature Trends

Figure 2a shows the joint relative frequency distribution of all available overlapping 30-year TTT *Ordinary Least Squares* (OLS) trend estimates across the 100 MPI-ESM realizations as a function of start year and trend size (CMIP5 counterparts shown in the SI Figure S1a). Observed and simulated trends show good agreement for the first two thirds of the series. For some observational estimates, trends with start years between 1965 and 1970 show higher warming than simulated, most likely due to the strong El Niño event coinciding with the 1983 El Chichón eruption. Simulated trends with start years after 1978 increasingly diverge from observations, with models simulating larger trends than those observed. Particularly, the period 1979–2008—identified in the literature as the *satellite era*, the most-often studied period—shows the lowest 30-year OLS trend in tropical upper tropospheric warming for the whole radiosonde record.

However, conclusions regarding differences between observed and simulated 30-year warming trends depend on the trend calculation method. In Figure 2b, we replicate our analysis of the L2-norm OLS linear regression (Figure 2a), but for L1-norm *Least Absolute Deviation* (LAD) linear regression [*Bloomfield and Steiger*, 1980; *Koenker*, 2005]. The LAD linear regression is a robust trend estimator that is less sensitive to outliers

in the data, such as strong Niño or Niña events or volcanic eruptions, while maintaining more of the interannual internal variability signal than the OLS estimator. Observed 30-year TTT trends calculated with LAD agree better with simulations over the recent decades, lying generally within the simulated ensemble spread. However, observed 30-year TTT trends calculated with both LAD and OLS estimators present a wide spread of roughly 0.2°C per decade for different radiosonde compilations, with IUKv2 estimating the highest observed trends, while model spread is roughly 0.4°C per decade. This broad observational spread indicates substantive remaining observational uncertainties in the TTT radiosonde record.

For 30-year trends of TTT-minus-ST, the difference between annual tropical upper troposphere temperature anomalies minus surface temperature anomalies, the range of observational estimates is roughly of the same magnitude as the 100 realization large-ensemble spread, about 0.20°C per decade (Figure 2c). Simulated 30-year trends of TTT-minus-ST show good agreement with observations for start years before 1972, but discrepancies appear for trends starting after that point. These discrepancies are more marked than for TTT trends and are only marginally smaller for LAD trend estimates (figure not shown), suggesting potential inconsistencies in tropical tropospheric warming amplification. However, due to the relatively short observational record, overlapping 30-year trends for the period 1958–2014 are not independent from each other. This dependence could lead to an underestimation of the observed long-term internal variability.

Marotzke and Forster [2015] find that different mechanisms drive the ensemble spread in globally averaged surface temperature trends for different trend lengths. Such differences between time scales will presumably also apply for tropospheric temperatures. We replicate our analysis for 15-year trends (Figure 3; for CMIP5 data in SI Figure S1b) and find relatively good agreement between observations and simulations. Overall, 15-year TTT trends are more influenced by interannual variability. Observational trend estimates are distributed over the whole simulated ensemble width but rarely lie outside the limits of the ensemble spread. In contrast to the 30-year TTT trends, there are only minor differences between TTT 15-year trends based on LAD or OLS trend estimators. Observed LAD 15-year trends with start years from 1962 to 1963 and from 1985 to 1988 fall outside the large-ensemble spread for most observational estimates (Figure 3b).

For the joint relative frequency distribution of 15-year TTT-minus-ST simulated trends, we find good agreement with observational trend estimates, with a reduction of the spread in simulated trends with respect to TTT trends (Figure 3c). Furthermore, we find a broad observational spread, of about 0.30°C per decade, that is about half of the large-ensemble spread. Observed TTT-minus-ST trends with start years between 1986 and 1991, covering the period 1986–2005, lie markedly outside the ensemble spread. For this period, we find lower than expected trends for both the tropical average and for trends at single radiosonde stations (figure not shown). This localized discrepancy period, which overlaps with the period of discrepancy in 30-year trends, stands in contrast to the good agreement found over the rest of the record and suggests period-dependent mechanisms, like remaining observational biases or an overestimated response to the Pinatubo eruption, rather than systematic flaws in model performance as the main cause for remaining upper tropospheric discrepancies.

To further evaluate these remaining discrepancies, we study the amplification of tropical surface warming by considering upper troposphere temperatures against simultaneous surface temperatures (Figure 4). Simulated annual temperatures show an amplification behavior consistent with observations as well as with theoretical expectations. MPI-ESM presents an average amplification factor of about 1.9—each 1°C of surface warming roughly amplifies to 1.9°C of upper tropospheric warming. For 15-and 30-year trends, the MPI-ESM amplification factor decreases to about 1.5 and 1.2, respectively. However, there is some variability in these factors that is larger for trends calculated over longer periods. For different realizations from the large ensemble, amplification factors for 30-year trends vary from 0.4 to 2.1, while they range from 1.8 to 2.2 for annual temperatures. These values show good agreement with the average amplification factor of the CMIP5 ensemble of about 1.8 [*Mitchell et al.*, 2013].

We also find that TTT trends present an offset to surface trends that is again larger for trends calculated over longer periods. For 30-year trends, the MPI-ESM shows upper tropospheric trends of around 0.13°C per decade on average for near to zero surface trends (Figure 4c). A physical explanation for this offset is that the surface warming in the neighboring latitude bands 20° to 30° North/South as well as in grid cells without radiosonde data may influence the tropical temperatures aloft due to the strong homogeneity of the tropical troposphere. Surface trends have been mostly positive during the observational period; therefore, the influence of neighboring regions could lead to positive upper tropospheric trends even for zero surface

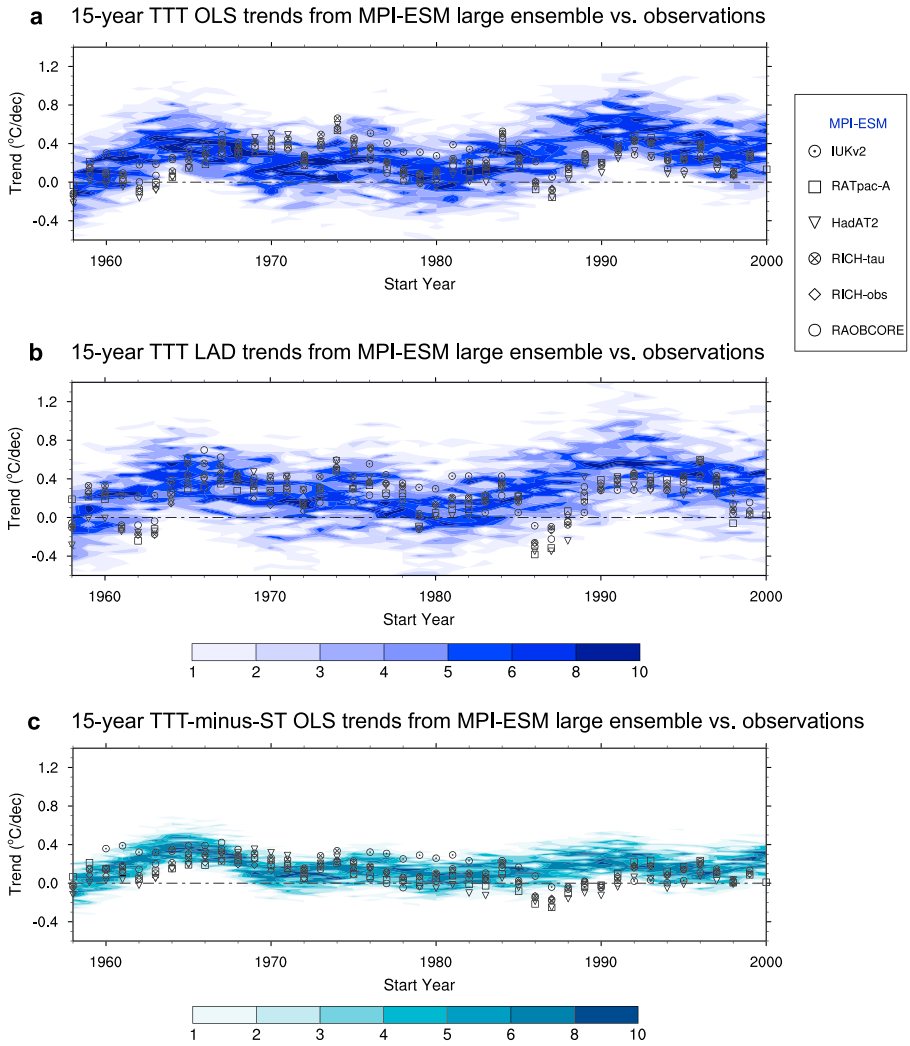

Figure 3. Simulated and observed 15-year trends. (a) Joint relative frequency distribution of MPI-ESM-simulated TTT *OLS* trend estimates compared to radiosonde estimates (black markers). Start years are separated by calendar year; bin size is 0.04°C per decade; and the color bar represents the ensemble member density, in percentage. (b) Joint relative frequency distribution of MPI-ESM-simulated TTT *LAD* trend estimates compared to radiosonde estimates, as in Figure 3a. Start years are separated by calendar year; bin size is 0.04°C per decade. (c) Joint relative frequency distribution of MPI-ESM-simulated TTT-minus-ST *OLS* trend estimates compared to radiosonde observational estimates. Bin size is 0.02°C per decade. For all radiosonde observations, the difference is shown against surface temperature values from the HadCRUT4.5 data set.

trends in the considered grid cells. For preindustrial conditions, we find that the effect of internal variability through the compensating influence of the neighboring regions practically eliminates this offset (SI Figure S2). MPI-ESM-simulated 30-year trends for preindustrial conditions also present an amplification factor of about 1.9, indicating that the aforementioned influence of neighboring regions may also be the cause of the increased range in amplification factors with trend length in the period 1958–2014.

For observations, annual estimates of TTT-versus_ST occur always within the MPI-ESM ensemble spread and present an amplification factor of about 1.8, which is within the internal variability range in simulated values (Figure 4a). For 15-year trends, observational estimates lie mostly within the large-ensemble spread, with the exception of trends with start years between 1986 and 1990, that lie marginally outside the ensemble spread for some compilations. Trend estimates with start years from 1986 to 1990, calculated over periods covering the 1986–2005 interval, show slightly less amplification than MPI-ESM. Some observational compilations present dampening of surface warming during this interval, with smaller warming trends in the upper troposphere than at the surface, an effect that is not present for trends over other periods (Figure 4b).

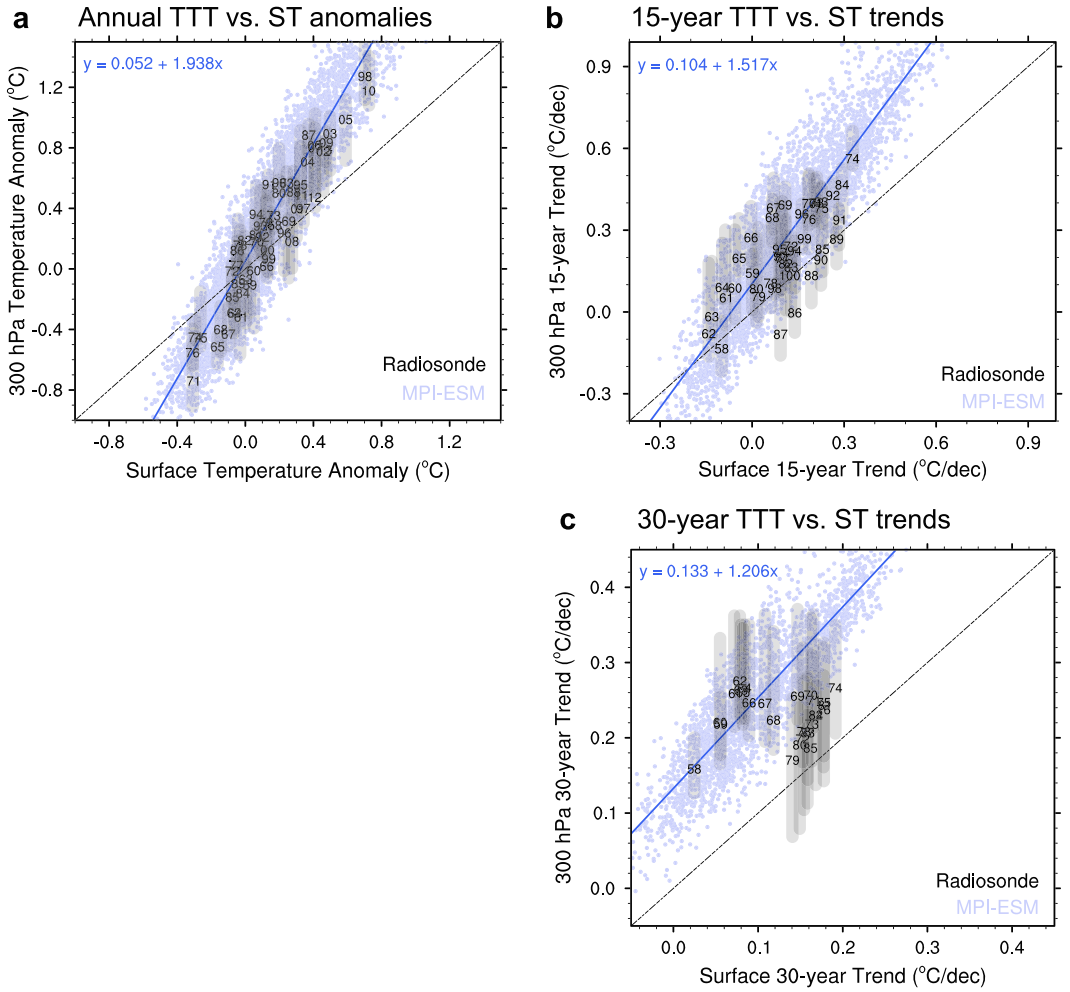

Figure 4. Simulated and observed TTT versus ST. (a) MPI-ESM-simulated and radiosonde-observed annual tropical tropospheric temperature anomalies versus MPI-ESM-simulated and HadCRUT4.5 surface temperature anomalies for the period 1958–2014. Radiosonde estimates are represented by year (black), arranged at the median of all observational estimates; gray bars indicate the observational spread. MPI-ESM estimates are represented by blue dots, and the blue line represents the OLS-fitted regression for the MPI-ESM large ensemble. For MPI-ESM $r^2 \approx 0.88$, for observations $r^2 \approx 0.89$. The black dashed line shows the one-to-one relationship; for positive surface trends, values above the dashed line indicate tropospheric amplification. (b) MPI-ESM-simulated and radiosonde-observed tropical tropospheric temperature versus MPI-ESM-simulated and HadCRUT4.5 surface temperature 15-year OLS trend estimates for the period 1958–2014. Radiosonde estimates are represented by start year (black), arranged at the median of all observational trend estimates, as for Figure 4a. For MPI-ESM $r^2 \approx 0.80$, for observations $r^2 \approx 0.54$. (c) MPI-ESM-simulated and radiosonde-observed tropical tropospheric temperature versus MPI-ESM-simulated and HadCRUT4.5 surface temperature 30-year OLS trend estimates, as for Figure 4b. For MPI-ESM $r^2 \approx 0.75$, for observations $r^2 \approx 0$.

For 30-year trends, simulations and observations show good agreement for trends with start years before 1971. For start years after 1971, some observational estimates occur inside the ensemble spread (e.g., IUKv2), yet again, other fall outside the ensemble spread and present dampening of surface warming on the upper troposphere (Figure 4c).

However, the validity of a simple proportionality law to quantify the amplification of surface warming aloft during recent decades appears to depend on the time scale considered. On annual time scales, there is a strong linear correlation between surface and upper troposphere temperature anomalies for both observational and model estimates. This linear correlation decreases with trend length for the period 1958–2014, as the variability range in the amplification factor increases. These results highlight the limitations of the concept of an amplification factor, particularly for longer trend lengths and when surface trends are close to zero.

The simulated amplification of tropical surface warming in the upper troposphere in MPI-ESM shows good agreement with both observations and theory on annual time scales, as well as for the majority of the trend record. Some observational compilations show less amplification or even dampening of surface warming aloft for trends calculated over periods covering the 1986–2005 interval. However, the most recently developed compilations present amplification regimes within the simulated range across all the time scales studied once internal variability is considered. Our results suggest that remaining upper tropospheric discrepancies arise most likely through period-dependent mechanisms or biases rather than systematic flaws in model performance.

4. Summary and Conclusions

We analyze the ability of two state-of-the-art model ensembles to simulate the observed variability in tropical upper troposphere temperatures (TTTs) as well as TTT trends over the period 1958–2014, to revisit the extensively discussed discrepancies between observed and simulated tropical tropospheric warming trends over the last decades. To our knowledge, this is the first study to use a 100 member single-model ensemble for this purpose, allowing us to characterize the role of internal variability in a transient setting.

Once simulated internal variability is considered, both the CMIP5 and the MPI-ESM ensembles adequately represent the radiosonde-observed annual large-scale variability in TTT records for the period 1958–2014. Our results also indicate that the spread in the CMIP5 ensemble can be considered an adequate representation of model-simulated internal variability for tropical upper tropospheric temperatures.

For 30-year TTT trends, we find good agreement between observational and simulated estimates with start years before 1978, with discrepancies between simulations and observations for OLS trend estimates with start years thereafter. This discrepancy is reconciled when the LAD estimator is used. We also find little discrepancies for 15-year TTT trends, estimated with either method. There is, however, a broad spread in observational estimates from different radiosonde compilations. This broad spread indicates remaining observational uncertainties, which are more relevant over 30-year trends and, particularly, when trends on TTT-minus-ST are considered. Our analysis of 15-year TTT-minus-ST trends shows good agreement between observations and simulations for the majority of the radiosonde record, with the exception of a few estimates with start years between 1986 and 1991. For 30-year TTT-minus-ST trends, we find good agreement for the first half of the record, but discrepancies appear when trends are calculated over periods including the 1986–2005 interval. Due to the longer periods covered by 30-year trends, these discrepancies span over half of the record, instead of over approximately 15% as for the 15-year trends.

For the amplification of surface warming in the upper troposphere, we find good agreement for all annual estimates, for all 15-year trends and for 30-year trends with start years before 1971. Observed trends calculated over periods covering the 1986–2005 interval show less tropospheric amplification than MPI-ESM for some observational compilations, while the most recently developed compilations show amplification factors within the simulated internal variability range across all time scales considered. Our results present, therefore, improved agreement between model simulations and observations, as well as with theoretical expectations, in comparison with previous studies based on earlier versions of radiosonde data sets [*Santer et al.*, 2005; *Mitchell et al.*, 2013]. However, our findings also highlight the limitations of a simple linear proportionality law to quantify the tropospheric amplification of surface warming for longer trend lengths or in case of surface trends close to zero.

Despite the continuous efforts made in achieving higher levels of homogeneity, calibration, and accuracy in observations, our results indicate that residual structural uncertainties in observational tropospheric temperature estimates remain large and that these remaining observational uncertainties contribute to discrepancies between observed and simulated trends over the last decades. The future development of revised versions of observational data sets, correcting remaining observational uncertainties and inhomogeneities, may continue to improve the agreement between observed and simulated multidecadal variability for tropical tropospheric temperatures.

We conclude that GCMs adequately simulate the internal variability in tropospheric warming amplification responses across different time scales over most of the available record. We find, however, remaining discrepancies for trends calculated over periods including the years 1986 to 2005, particularly when considering trends in tropospheric amplification. The good agreement between simulated and observed tropospheric

warming for all annual estimates, all 15-year trends, and most of the 30-year trend record—once the role of internal variability is taken into account—suggests that the remaining discrepancies arise mainly because of observational biases or other period-dependent mechanisms, rather than because of systematic errors in model performance.

Acknowledgments
This research was supported by the Max Planck Society for the Advancement of Science. We acknowledge Luis Kornblueh, Jürgen Kröger, and Michael Botzet for producing the historical and RCP4.5 MPI-ESM large ensemble simulations, and the Swiss National Computing Centre (CSCS) and the German Climate Computing Center (DKRZ) for providing the necessary computational resources. We acknowledge the World Climate Research Programme's Working Group on Coupled Modelling, which is responsible for CMIP, and we thank the climate modeling groups (listed in Table S2 of the supporting information) for producing and making available their model output. For CMIP the U.S. Department of Energy's Program for Climate Model Diagnosis and Intercomparison provides coordinating support and led development of software infrastructure in partnership with the Global Organization for Earth System Science Portals. We acknowledge the Earth System Grid Federation (ESGF) portal as well as the German Climate Computing Center (DKRZ) for facilitating data access. We would also like to acknowledge the groups that developed and facilitated the different observational compilations (listed in Table S1 of the supporting information). We thank Wolfgang Müller and two anonymous reviewers for their helpful comments on the manuscript. Scripts used in the analysis and other supporting information that may be useful in reproducing the authors work are archived by the Max Planck Institute for Meteorology and can be obtained by contacting publications@mpimet.mpg.de.

References

Annan, J. D., and J. C. Hargreaves (2010), Reliability of the CMIP3 ensemble, *Geophys. Res. Lett.*, *37*, L02703, doi:10.1029/2009GL041994.
Bentsen, M., et al. (2013), The Norwegian Earth System Model, NorESM1-M—Part 1: Description and basic evaluation of the physical climate, *Geosci. Model Dev.*, *6*, 687–720, doi:10.5194/gmd-6-687-2013.
Bi, D., et al. (2013), The ACCESS coupled model: Description, control climate and evaluation, *Aust. Meteorol. Oceanogr. J.*, *63*, 41–64.
Bloomfield, P., and W. Steiger (1980), Least absolute deviations curve-fitting, *SIAM J. Sci. Stat. Comput.*, *1*, 290–301.
Collins, W. J., et al. (2011), Development and evaluation of an Earth-System model—HadGEM2, *Geosci. Model Dev.*, *4*, 1051–1075, doi:10.5194/gmd-4-1051-2011.
Colman, R. A. (2001), On the vertical extent of atmospheric feedbacks, *Clim. Dyn.*, *17*, 391–405, doi:10.1007/s003820000111.
Delworth, T. L., et al. (2006), GFDL's CM2 global coupled climate models. Part I: Formulation and simulation characteristics, *J. Clim.*, *19*, 643–674, doi:10.1175/JCLI3629.1.
Donner, L. J., et al. (2011), The dynamical core, physical parameterizations, and basic simulation characteristics of the atmospheric component AM3 of the GFDL global coupled model CM3, *J. Clim.*, *24*, 3484–3519, doi:10.1175/2011JCLI3955.1.
Douglass, D. H., J. R. Christy, B. D. Pearson, and S. F. Singer (2007), A comparison of tropical temperature trends with model predictions, *Int. J. Climatol.*, *28*, 1693–1701, doi:10.1002/joc.1651.
Dufresne, J.-L., et al. (2013), Climate change projections using the IPSL-CM5 Earth system model: From CMIP3 to CMIP5, *Clim. Dyn.*, *40*, 2123–2165, doi:10.1007/s00382-012-1636-1.
Flato, G. J., 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, doi:10.1017/CBO9781107415324.020.
Free, M., D. J. Seidel, J. K. Angell, J. R. Lanzante, and T. C. Peterson (2005), Radiosonde atmospheric temperature products for assessing climate (RATPAC): A new data set of large area anomaly time series, *J. Geophys. Res.*, *110*, D22101, doi:10.1029/2005JD006169.
Fu, Q., C. M. Johanson, S. G. Warren, and D. J. Seidel (2004), Contribution of stratospheric cooling to satellite-inferred tropospheric temperature trends, *Nature*, *429*, 55–58, doi:10.1038/nature02524.
Fu, Q., S. Manabe, and C. M. Johanson (2011), On the warming in the tropical upper troposphere: Models versus observations, *Geophys. Res. Lett.*, *38*, L15704, doi:10.1029/2011GL048101.
Fyfe, J. C., N. P. Gillett, and F. W. Zwiers (2013), Overestimated global warming over the past 20 years, *Nat. Clim. Change*, *3*, 767–769, doi:10.1038/nclimate1972.
Giorgetta, M. A., et al. (2013), Climate and carbon cycle changes from 1850 to 2100 in MPI-ESM simulations for the Coupled Model Intercomparison Project phase 5, *J. Adv. Model. Earth Syst.*, *5*, 572–597, doi:10.1002/jame.20038.
Haimberger, L. (2007), Homogenization of radiosonde temperature time series using innovation statistics, *J. Clim.*, *20*, 1377–1403, doi:10.1175/JCLI4050.1.
Haimberger, L., C. Tavolato, and S. Sperka (2012), Homogenization of the global radiosonde temperature dataset through combined comparison with reanalysis background series and neighboring stations, *J. Clim.*, *25*, 8108–8131, doi:10.1175/JCLI-D-11-00668.1.
Hamill, T. H. (2001), Interpretation of rank histograms for verifying ensemble forecasts, *Mon. Weather Rev.*, *129*, 550–560, doi:10.1175/1520-0493(2001)129<0550:IORHFV>2.0.CO;2.
Hazeleger, W., C. Severijns, T. Semmler, S. Stefanescu, S. Yang, X. Wang, K. Wyser, E. Dutra, and J. M. Baldasano (2010), EC-EARTH: A seamless Earth-System prediction approach in action, *Bull. Am. Meteorol. Soc.*, *91*, 1357–1363, doi:10.1175/2010BAMS2877.1.
Hourdin, F., et al. (2013), LMDZ5B: The atmospheric component of the IPSL climate model with revisited parameterizations for clouds and convection, *Clim. Dyn.*, *40*, 2193–2222, doi:10.1007/s00382-012-1343-y.
Hurrell, J. W., et al. (2012), The community Earth system model: A framework for collaborative research, *Bull. Am. Meteorol. Soc.*, *94*, 1339–1360, doi:10.1175/BAMS-D-12-00121.1.
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. 571–657, Cambridge Univ. Press, Cambridge, U. K., and New York.
Ji, D., et al. (2014), Description and basic evaluation of Beijing Normal University Earth system model (BNU-ESM) version 1, *Geosci. Model Dev. Discuss.*, *7*, 2039–2064, doi:10.5194/gmdd-7-1601-2014.
Jones, C. D., et al. (2011), The HadGEM2-ES implementation of CMIP5 centennial simulations, *Geosci. Model Dev.*, *4*, 543–570, doi:10.5194/gmd-4-543-2011.
Knight, J., J. Kennedy, C. K. Folland, G. Harris, G. S. Jones, and M. D. Palmer (2010), Do global temperature trends over the last decade falsify climate predictions?, *Bull. Am. Meteorol. Soc.*, *90*, S22–S23.
Koenker, R. (2005), *Quantile Regression, Econometric Soc. Monogr. Ser.*, Cambridge Univ. Press, Cambridge, U. K.
Lanzante, J. R., S. A. Klein, and D. J. Seidel (2003), Temporal homogenization of monthly radiosonde temperature data. Part II: Trends, sensitivities and MSU comparison, *J. Clim.*, *16*, 241–262, doi:10.1175/1520-0442(2003)016<0241:THOMRT>2.0.CO;2.
Li, L., et al. (2013), The flexible global ocean-atmosphere-land system model, grid-point version 2: FGOALS-g2, *Adv. Atmos. Sci.*, *30*, 543–560, doi:10.1007/s00376-012-2140-6.
Liebmann, B., R. M. Dole, C. Jones, I. Blade, and D. Allured (2010), Influence of choice of time period on global surface temperature trend estimates, *Bull. Am. Meteorol. Soc.*, *91*, 1485–1491, doi:10.1175/2010BAMS3030.1.
Manabe, S., and R. J. Stouffer (1980), Sensitivity of a global climate model to an increase of CO^2 concentration in the atmosphere, *J. Geophys. Res.*, *85*, 5529–5554, doi:10.1029/JC085iC10p05529.
Marotzke, J., and P. M. Forster (2015), Forcing, feedback and internal variability in global temperature trends, *Nature*, *517*, 565–570, doi:10.1038/nature14117.
McKitrick, R., S. McIntyre, and C. Herman (2010), Panel and multivariate methods for tests of trend equivalence in climate data series, *Atmos. Sci. Lett.*, *11*, 270–277, doi:10.1002/asl.290.
Meehl, G. A., et al. (2012), Climate system response to external forcings and climate change projections in CCSM4, *J. Clim.*, *25*, 3661–3683, doi:10.1175/JCLI-D-11-00240.1.

Mitchell, D. M., P. W. Thorne, P. A. Stott, and L. J. Gray (2013), Revisiting the controversial issue of tropical tropospheric temperature trends, *Geophys. Res. Lett.*, *40*, 2801–2806, doi:10.1002/grl.50465.
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.
Osborn, T. J., and P. D. Jones (2014), The CRUTEM4 land-surface air temperature data set: Construction, previous versions and dissemination via Google Earth, *Earth Syst. Sci. Data*, *6*(1), 61–68, doi:10.5194/essd-6-61-2014.
Po-Chedley, S., and Q. Fu (2012), Discrepancies in tropical upper tropospheric warming between atmospheric circulation models and satellites, *Environ. Res. Lett.*, *7*, 044018, doi:10.1088/1748-9326/7/4/044018.
Po-Chedley, S., T. J. Thorsen, and Q. Fu (2015), Removing diurnal cycle contamination in satellite-derived tropospheric temperatures: Understanding tropical tropospheric trend discrepancies, *J. Clim.*, *28*, 2274–2290, doi:10.1175/JCLI-D-13-00767.1.
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*, *4*, 835–840, doi:10.1038/nclimate2310.
Rotstayn, L. D., M. A. Collier, M. R. Dix, Y. Feng, H. B. Gordon, S. P. O'Farrell, I. N. Smith, and J. Syktus (2010), Improved simulation of australian climate and ENSO-related rainfall variability in a global climate model with an interactive aerosol treatment, *Int. J. Climatol.*, *30*, 1067–1088, doi:10.1002/joc.1952.
Sakamoto, T. T., et al. (2012), MIROC4h—A new high-resolution atmosphere-ocean coupled general circulation model, *J. Meteorol. Soc. Jpn. Ser. II*, *90*, 325–359, doi:10.2151/jmsj.2012-301.
Santer, B. D., T. M. L. Wigley, C. Doutriaux, J. S. Boyle, J. E. Hansen, P. D. Jones, G. A. Meehl, E. Roeckner, S. Sengupta, and K. E. Taylor (2001), Accounting for the effects of volcanoes and ENSO in comparisons of modeled and observed temperature trends, *J. Geophys. Res.*, *106*, 28,033–28,059, doi:10.1029/2000JD000189.
Santer, B. D., et al. (2005), Amplification of surface temperature trends and variability in the tropical atmosphere, *Science*, *309*, 1551–1556, doi:10.1126/science.1114867.
Santer, B. D., et al. (2008), Consistency of modelled and observed temperature trends in the tropical troposphere, *Int. J. Climatol.*, *28*, 1703–1722, doi:10.1002/joc.1756.
Santer, B. D., et al. (2014), Volcanic contribution to decadal changes in tropospheric temperature, *Nat. Geosci.*, *7*, 185–189, doi:10.1038/NGEO2098.
Santer, B. D., et al. (2016), Comparing tropospheric warming in climate models and satellite data, *J. Clim.*, *30*(1), 373–392, doi:10.1175/JCLI-D-16-0333.1.
Schmidt, G. A., et al. (2014), Configuration and assessment of the GISS ModelE2 contributions to the CMIP5 archive, *J. Adv. Model. Earth Syst.*, *6*, 141–184, doi:10.1002/2013MS000265.
Scoccimarro, E., S. Gualdi, A. Bellucci, A. Sanna, P. G. Fogli, E. Manzini, P. Vichi, M. Oddo, and A. Navarra (2011), Effects of tropical cyclones on ocean heat transport in a high resolution coupled general circulation model, *J. Clim.*, *24*, 4368–4384, doi:10.1175/2011JCLI4104.1.
Seidel, D. J., M. Free, and J. S. Wang (2012), Reexamining the warming in the tropical upper troposphere: Models versus radiosonde observations, *Geophys. Res. Lett.*, *39*, L22701, doi:10.1029/2012GL053850.
Sherwood, S. C., and N. Nishant (2015), Atmospheric changes through 2012 as shown by iteratively homogenized radiosonde temperature and wind data (IUKv2), *Environ. Res. Lett.*, *10*, 054007, doi:10.1088/1748-9326/10/5/054007.
Sherwood, S. C., J. R. Lanzante, and C. L. Meyer (2005), Radiosonde daytime biases and late-20th century warming, *Science*, *309*, 1556–1559, doi:10.1126/science.1115640.
Sherwood, S. C., C. L. Meyer, R. J. Allen, and H. A. Titchner (2008), Robust tropospheric warming revealed by iteratively homogenized radiosonde data, *J. Clim.*, *21*, 5336–5352, doi:10.1175/2008JCLI2320.1.
Smith, D. M., R. Eade, N. J. Dunstone, D. Fereday, J. M. Murphy, H. Pohlmann, and A. A. Scaife (2010), Skilful multi-year predictions of atlantic hurricane frequency, *Nat. Geosci.*, *3*, 846–849, doi:10.1038/ngeo1004.
Solomon, S., J. S. Daniel, R. R. Neely, J. P. Vernier, E. G. Dutton, and L. W. Thomason (2011), The persistently variable "background" stratospheric aerosol layer and global climate change, *Science*, *333*, 866–870, doi:10.1126/science.1206027.
Taylor, K. E., R. J. Stouffer, and G. A. Meehl (2012), An overview of CMIP5 and the experiment design, *Bull. Am. Meteorol. Soc.*, *93*, 485–498, doi:10.1175/BAMS-D-11-00094.1.
Thorne, P. W., D. E. Parker, S. F. B. Tett, P. D. Jones, M. McCarthy, H. Coleman, and P. Brohan (2005), Revisiting radiosonde upper-air temperatures from 1958 to 2002, *J. Geophys. Res.*, *110*, D18105, doi:10.1029/2004JD005753.
Thorne, P. W., D. E. Parker, B. D. Santer, M. P. McCarthy, D. M. H. Sexton, M. J. Webb, J. M. Murphy, M. Collins, H. A. Titchner, and G. S. Jones (2007), Tropical vertical temperature trends: A real discrepancy?, *Geophys. Res. Lett.*, *34*, L16702, doi:10.1029/2007GL029875.
Thorne, P. W., et al. (2011a), A quantification of uncertainties in historical tropical tropospheric temperature trends from radiosondes, *J. Geophys. Res.*, *116*, D12116, doi:10.1029/2010JD015487.
Thorne, P. W., J. R. Lanzante, T. C. Peterson, D. J. Seidel, and K. P. Shine (2011b), Tropospheric temperature trends: History of an ongoing controversy, *Wiley Interdiscip. Rev. Clim. Change*, *2*, 66–88, doi:10.1002/wcc.80.
Titchner, H. A., P. W. Thorne, M. P. McCarthy, S. F. B. Tett, L. Haimberger, and D. E. Parker (2009), Critically reassessing tropospheric temperature trends from radiosondes using realistic validation experiments, *J. Clim.*, *22*, 465–485, doi:10.1175/2008JCLI2419.1.
Voldoire, A., et al. (2013), The CNRM-CM5.1 global climate model: Description and basic evaluation, *Clim. Dyn.*, *40*, 2091–2121, doi:10.1007/s00382-011-1259-y.
Volodin, E. M., N. A. Dianskii, and A. V. Gusev (2010), Simulating present-day climate with the INMCM4.0 coupled model of the atmospheric and oceanic general circulations, *Izv. Atmos. Oceanic Phys.*, *46*, 414–431, doi:10.1134/S000143381004002X.
von Salzen, K., et al. (2013), The Canadian Fourth Generation Atmospheric Global Climate Model (CanAM4). Part I: Representation of physical processes, *Atmos. Ocean*, *51*, 104–125, doi:10.1080/07055900.2012.755610.
Wang, S., X. Wen, Y. Luo, Z. Zhao, and J. Huang (2010), Does the global warming pause in the last decade: 1999–2008, *Adv. Clim. Change Res.*, *1*, 49–54, doi:10.3724/SP.J.1248.2010.00049.
Watanabe, M., et al. (2010), Improved climate simulation by MIROC5: Mean states, variability, and climate sensitivity, *J. Clim.*, *23*, 6312–6335, doi:10.1175/2010JCLI3679.1.
Watanabe, S., et al. (2011), MIROC-ESM 2010: Model description and basic results of CMIP5-20c3m experiments, *Geosci. Model Dev.*, *4*, 845–872, doi:10.5194/gmd-4-845-2011.
Xin, X., T. Wu, J. Li, Z. Wang, W. Li, and F. Wu (2013), How well does BCC CSM1.1 reproduce the 20th century climate change over China?, *Atmos. Oceanic Sci. Lett.*, *6*, 21–26.
Yukimoto, S., et al. (2012), A new global climate model of the meteorological research institute: MRI-CGCM3-Model description and basic performance, *J. Meteorol. Soc. Jpn. Ser. II*, *90A*, 23–64, doi:10.2151/jmsj.2012-A02.