University of Reading

Historical simulations with HadGEM3-GC3.1 for CMIP6

Article

Published Version

Creative Commons: Attribution 4.0 (CC-BY)

Open Access

Andrews, M. B., Ridley, J. K., Wood, R. A., Andrews, T., Blockley, E. W., Booth, B., Burke, E., Dittus, A. J. ORCID: https://orcid.org/0000-0001-9598-6869, Florek, P., Gray, L. J., Haddad, S., Hardiman, S. C., Hermanson, L., Hodson, D. ORCID: https://orcid.org/0000-0001-7159-6700, Hogan, E., Jones, G. S., Knight, J. R., Kuhlbrodt, T. ORCID: https://orcid.org/0000-0003-2328-6729, Misios, S., Mizielinski, M. S., Ringer, M. A., Robson, J. ORCID: https://orcid.org/0000-0002-3467-018X and Sutton, R. T. ORCID: https://orcid.org/0000-0001-8345-8583 (2020)
Historical simulations with HadGEM3-GC3.1 for CMIP6. Journal of Advances in Modeling Earth Systems, 12 (6). e2019MS001995. ISSN 1942-2466 doi: 10.1029/2019MS001995 Available at https://centaur.reading.ac.uk/91172/

It is advisable to refer to the publisher's version if you intend to cite from the work. See Guidance on citing.

To link to this article DOI: http://dx.doi.org/10.1029/2019MS001995

Publisher: American Geophysical Union

University of Reading

All outputs in CentAUR are protected by Intellectual Property Rights law, including copyright law. Copyright and IPR is retained by the creators or other copyright holders. Terms and conditions for use of this material are defined in the End User Agreement.

www.reading.ac.uk/centaur

CentAUR

Central Archive at the University of Reading

Reading's research outputs online

JAMES | Journal of Advances in Modeling Earth Systems

RESEARCH ARTICLE
10.1029/2019MS001995

Special Section:
The UK Earth System Models for CMIP6

Historical Simulations With HadGEM3-GC3.1 for CMIP6

Martin B. Andrews[1], Jeff K. Ridley[1], Richard A. Wood[1], Timothy Andrews[1], Edward W. Blockley[1], Ben Booth[1], Eleanor Burke[1], Andrea J. Dittus[2], Piotr Florek[1], Lesley J. Gray[3], Stephen Haddad[1], Steven C. Hardiman[1], Leon Hermanson[1], Dan Hodson[2], Emma Hogan[1], Gareth S. Jones[1], Jeff R. Knight[1], Till Kuhlbrodt[2], Stergios Misios[3], Matthew S. Mizielinski[1], Mark A. Ringer[1], Jon Robson[2], and Rowan T. Sutton[2]

[1]Met Office Hadley Centre, Exeter, UK, [2]NCAS, Department of Meteorology, University of Reading, Reading, UK, [3]NCAS, Department of Physics, University of Oxford, Oxford, UK

Abstract We describe and evaluate historical simulations which use the third Hadley Centre Global Environment Model in the Global Coupled configuration 3.1 (HadGEM3-GC3.1) model and which form part of the UK's contribution to the sixth Coupled Model Intercomparison Project, CMIP6. These simulations, run at two resolutions, respond to historically evolving forcings such as greenhouse gases, aerosols, solar irradiance, volcanic aerosols, land use, and ozone concentrations. We assess the response of the simulations to these historical forcings and compare against the observational record. This includes the evolution of global mean surface temperature, ocean heat content, sea ice extent, ice sheet mass balance, permafrost extent, snow cover, North Atlantic sea surface temperature and circulation, and decadal precipitation. We find that the simulated time evolution of global mean surface temperature broadly follows the observed record but with important quantitative differences which we find are most likely attributable to strong effective radiative forcing from anthropogenic aerosols and a weak pattern of sea surface temperature response in the low to middle latitudes to volcanic eruptions. We also find evidence that anthropogenic aerosol forcings play a role in driving the Atlantic Multidecadal Variability and the Atlantic Meridional Overturning Circulation, which are key features of the North Atlantic ocean. Overall, the model historical simulations show many features in common with the observed record over the period 1850–2014 and so provide a basis for future in-depth study of recent climate change.

Plain Language Summary Historical simulations, which successfully reproduce features of the observed climate from the end of the preindustrial period to the near-present day, contribute to our understanding of the underlying mechanisms that drive or influence climate variability and climate change. The historical simulations described in this paper use the third Hadley Centre Global Environment Model in the Global Coupled configuration 3.1 model. These simulations form part of the UK's contribution to the sixth Coupled Model Intercomparison Project. We assess various aspects of the historical climate system in our simulations against observations. This includes the evolution of global mean surface temperature, ocean heat content, sea ice extent, ice sheet mass balance, permafrost extent, snow cover, North Atlantic sea surface temperature and circulation, and decadal precipitation. The key findings include (a) that the model global mean surface temperature broadly follows the observed record, with a few quantitative differences, and (b) that the ocean circulation and sea surface temperatures of the North Atlantic are likely influenced by historical forcings. In general, the simulations respond to historically evolving influences in a similar manner to the observed world. Therefore, these simulations contribute, as part of the wider CMIP6 multimodel effort, to the understanding of the causes of observed climate change since 1850.

Key Points:
- We describe and evaluate the UK's CMIP6 historical simulations
- We identify drivers of the modeled global temperature evolution and compare with the observed record
- We find that anthropogenic aerosol forcings influence the simulation of North Atlantic ocean variability

Correspondence to:
M. B. Andrews,
martin.andrews@metoffice.gov.uk

Citation:
Andrews, M. B., Ridley, J. K., Wood, R. A., Andrews, T., Blockley, E. W., Booth, B., et al. (2020). Historical simulations with HadGEM3-GC3.1 for CMIP6. *Journal of Advances in Modeling Earth Systems*, *12*, e2019MS001995. https://doi.org/10.1029/2019MS001995

Received 19 DEC 2019
Accepted 4 MAY 2020
Accepted article online 11 MAY 2020

©2020. The Authors.
This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

1. Introduction

This paper describes and evaluates the historical simulations which form part of the UK's contribution to the sixth Coupled Model Intercomparison Project (CMIP6 Eyring et al., 2016), organized under the auspices of the World Climate Research Programme's Working Group on Coupled Modeling. These historical simulations, combined with the Diagnostic, Evaluation and Characterization of Klima (DECK) simulations described elsewhere in this special issue, form the UK's entry card to model participation in CMIP6.

The assessment of historical simulations contributes to the understanding of recent climate variability and climate change and forms part of the continuum toward the understanding of future climate change. The simulations also provide the opportunity to evaluate the ability of models to simulate climate variability and century timescale trends. The UK's historical simulations described in this paper will be assessed alongside those from other climate centers as part of the CMIP6 multimodel analysis.

The historical period (1850–2014), from approximately the end of the industrial revolution through to the near-present day, represents a time span over which the anthropogenic influence on greenhouse gases, aerosols, and land use has increased. This period includes a number of major tropical volcanic eruptions and solar variability, which can have a climatic influence on annual-to-decadal timescales. The observational record provides evidence of the combined impact of natural and anthropogenic forcings, the caveats being that these records are often short and globally incomplete and that they also include the stochastic influence of internal variability of the climate system. Despite these limitations, the approach adopted in this paper is to benchmark the historical experiments against the observational record to ascertain whether the simulations can recreate observed phenomena. This evaluation can also provide insight and understanding of how the Earth's climate has evolved over the historical period.

We focus on phenomena considered important for the characterization of coupled atmosphere-ocean historical experiments. These include the evolution of global mean surface air temperature, decadal precipitation, and ocean heat content (OHC), which are analyzed in section 2 and sea ice extent, ice sheet mass balance, permafrost extent, and snow cover in section 3. We consider the North Atlantic climate variability in terms of long-term climate modes, the overturning circulation and heat content in section 4. We also consider the transient response to major tropical volcanic eruptions and the solar cycle (SC) in sections 5 and 6, respectively. In section 7 we present an initial analysis of the mechanisms driving the modeled evolution of global mean surface temperature and its differences from the observed record. The assessment for any one of these phenomena could provide enough material for an entire standalone paper, hence the need for brevity herein. The objective of this paper is to demonstrate that the historical simulations show many features in common with the observed record and, therefore, will provide a useful contribution to the multimodel analysis to be performed as part of CMIP6.

These historical simulations also provide initial conditions for CMIP6 “future” scenario simulations, having been “spun-up” with 165 years of time-varying historical forcings.

1.1. Model Configuration

The simulations are run using the third Hadley Centre Global Environment Model in the Global Coupled configuration 3.1 (HadGEM3-GC3.1 Williams et al., 2017), the same model as used for the preindustrial control (*piControl*) simulations (Menary et al., 2018) from which the initial conditions are obtained. This physical climate model consists of coupled atmosphere, ocean, sea ice, and land models. It includes the global atmosphere-land configuration GA7/GL7.1 (Walters et al., 2019), the global ocean GO6 (Storkey et al., 2018) which uses the Nucleus for European Modeling of the Ocean (NEMO Madec et al., 2017) at vn3.6, and the sea ice model configuration GSI8.1 (Ridley et al., 2018) which uses the CICE5.1 model (Hunke et al., 2015). Coupling between the atmosphere and ocean models uses the OASIS-MCT coupler (Valcke et al., 2015). The two atmosphere/ocean resolutions used for the historical simulations are N96/ORCA1 and N216/ORCA025. These are referred to as GC31-LL and GC31-MM, respectively, and, apart from a few differences in parameter values associated with model resolution (Kuhlbrodt et al., 2018; Roberts et al., 2019), have a similar scientific configuration. The GC31-LL configuration (Kuhlbrodt et al., 2018) has a nominal atmospheric resolution of 135 km (N96) and an ocean resolution of 1° (ORCA1), with coupling every 3 hr. The GC31-MM configuration (Williams et al., 2017) has a nominal atmospheric resolution of 60 km (N216), and an ocean resolution of 0.25° (ORCA025), with coupling every hour. The atmosphere for both configurations has 85 model levels, up to 85 km, and the ocean has 75 levels.

1.2. Historical Forcings

The historical simulations employ CMIP6-defined forcings which are time varying and in some cases spatially varying. These forcings are briefly described below, and detailed, together with a description of how the forcings are adapted for HadGEM3-GC3.1, in Sellar et al. (2020).

Greenhouse gases (CO_2; CH_4; N_2O; CFCs; HFCs) are included in the model as global annual mean values. Primary emissions of black carbon and organic carbon aerosol (representing the effects of fossil fuel, biofuel,

and biomass burning), plus gas phase emissions of SO_2 (representing continuously degassing volcanoes, DMS, and monoterpene) are included in the model as monthly latitude-longitude fields. The aerosols generated by volcanic eruptions are represented in the model as monthly zonal mean optical parameters in the solar (shortwave) and terrestrial (longwave) spectrum on each model level.

The solar variability data set (Matthes et al., 2017), which includes the approximately 11-year SC, with longer-term variability, is included in the model as monthly mean values and varies across spectral bands. The historical three-dimensional monthly varying CMIP6 ozone concentrations are prescribed in the model. A potential mismatch between the tropopause height in the prescribed ozone and the varying tropopause height in the simulations required the development of an ozone remapping scheme which redistributes the CMIP6 ozone concentrations based on model tropopause evolution (Hardiman et al., 2019). The tropopause height in the historical simulations varies by only a small number of model levels as the climate warms toward the near-present day. The future scenario simulations, which branch from the end of the historical simulations, are more likely to be significantly impacted if the tropopause mismatch is unresolved. The ozone remapping scheme is used in the GC31-LL historical simulations. The limited time available to rerun the higher resolution, GC31-MM, historical simulations meant that they were run without the remapping scheme from 1850 to 1950 and with the remapping scheme effective from 1951 onward. The minimal impact of this midrun scientific upgrade is assessed in Appendix A1.

Historical monthly varying land use change, which covers nine land surface types (urban, bare soil, lake, ice, and five plant functional types), is included in the model.

1.3. Experiment Description

Two historical ensembles were run from 1 January 1850 to 31 December 2014, one ensemble per model resolution, GC31-LL and GC31-MM. Each ensemble consists of four members which may help mitigate the impact of model internal variability when analyzing the ensemble mean fields. The influence of resolution on the ability to simulate climate phenomena can be assessed by comparing the two ensembles.

1.4. Initialization

The historical experiments start on 1 January 1850 with initial conditions generated from the *piControl* experiments. The GC31-LL historical experiments used initial conditions from various points in the numerical integration of the GC31-LL *piControl* run, and, for the GC31-MM historical experiments, from the GC31-MM *piControl* run (Kuhlbrodt et al., 2018; Menary et al., 2018). Ideally, the ocean initial conditions for each experiment are chosen to maximize the possibility of ensemble members evolving independently. This reduces the possibility that significant responses to historical forcings will be confounded by long-term internal variability common across ensemble members. To this end, we decided to minimize the chance of experiments starting from similar oceanic states by considering the two leading modes of multidecadal sea surface temperature (SST) variability, the Interdecadal Pacific Oscillation (IPO; Power et al., 1999; Zhang et al., 1997) and the Atlantic Multidecadal Oscillation (AMO; Kerr, 2000), which naturally arise in the *piControl* runs (Menary et al., 2018). Both the IPO and AMO can have impacts on climate up to the global scale. We cannot prevent chance alignment of these modes later in the simulations, but we can try to ensure that this does not occur in the early stages, since the initial state of these modes can affect the simulation for several decades. The leading Empirical Orthogonal Function of the deseasonalized, detrended, and low-pass filtered Pacific basin SSTs, representing the IPO, for each of the *piControl* runs is shown in Figures 1a and 1c. The leading mode of variability for the North Atlantic basin, calculated as the area-weighted mean time series of the deseasonalized, detrended, and low-pass filtered North Atlantic SSTs, representing the AMO, is shown in Figures 1b and 1d. Low-pass filtering the data prior to calculation minimizes the influence of El Niño–Southern Oscillation. The monthly evolution of the IPO and AMO indices is plotted on phase-space diagrams in Figure 2. The phase-space diagrams are used to identify initial conditions which differ in terms of ocean state, differ in trajectory (not shown) through the phase-space, and which are separated by at least 20 years. The GC31-LL historical experiments were initialized with initial conditions from 1850, 1885, 1930, and 1970 of the GC31-LL *piControl* experiment and for the GC31-MM historical experiment from 1850, 1885, 1930, and 1960 of the GC31-MM *piControl* experiment. The selection of oceanic modes for GC31-LL initial conditions includes positive and negative IPO and AMO phases. For GC31-MM the initial conditions were chosen prior to the completion of the *piControl* experiment, resulting in IPO and AMO phases that appear less significant when considering the entire 500-year *piControl* experiment.

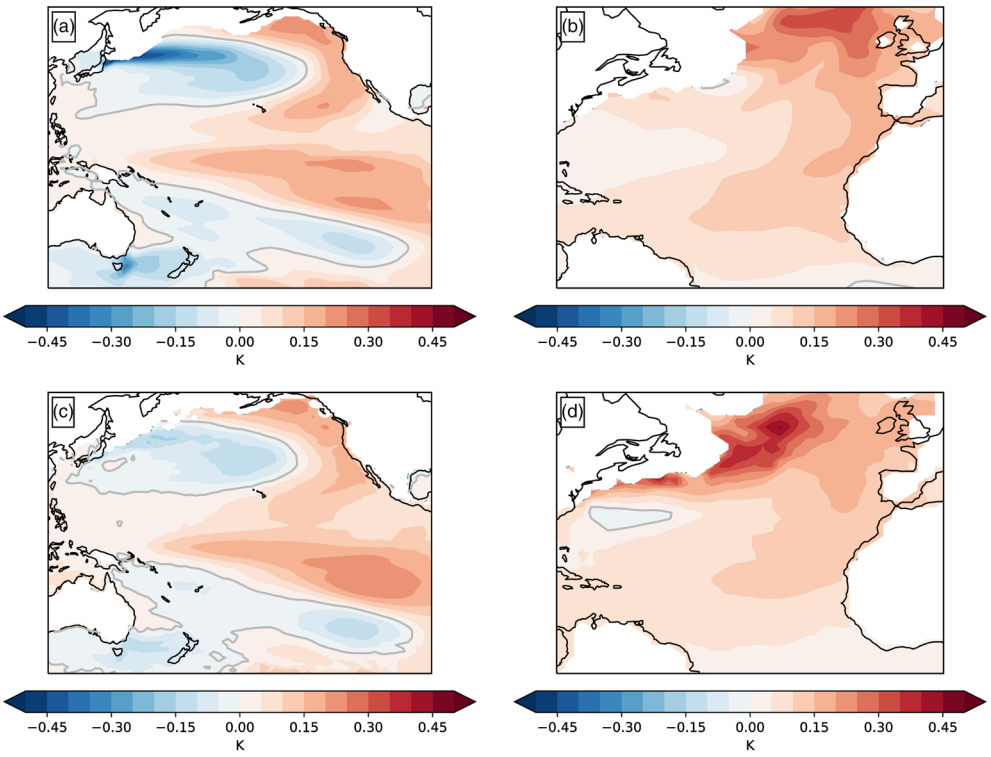

Figure 1. Ocean surface temperature modes internally generated by the 500-year *piControl* experiments. (a, b) The IPO and AMO patterns, respectively, for the GC31-LL *piControl* experiment. (c, d) The equivalent for the GC31-MM *piControl* experiment.

2. Temperature, OHC, and Precipitation

The response of the climate system to evolving historical forcings, in particular the combination of greenhouse gases, atmospheric aerosols and volcanic eruptions, is evident in the historical observational record of global mean surface temperature (Figure 3 Hartmann et al., 2013, Figure 2.20). To assess the ability of our historical experiments to reproduce features such as centennial temperature trends and rapid near-present-day warming, we examine the global mean surface air temperature evolution. The cooling immediately following volcanic eruptions is analyzed in section 5.

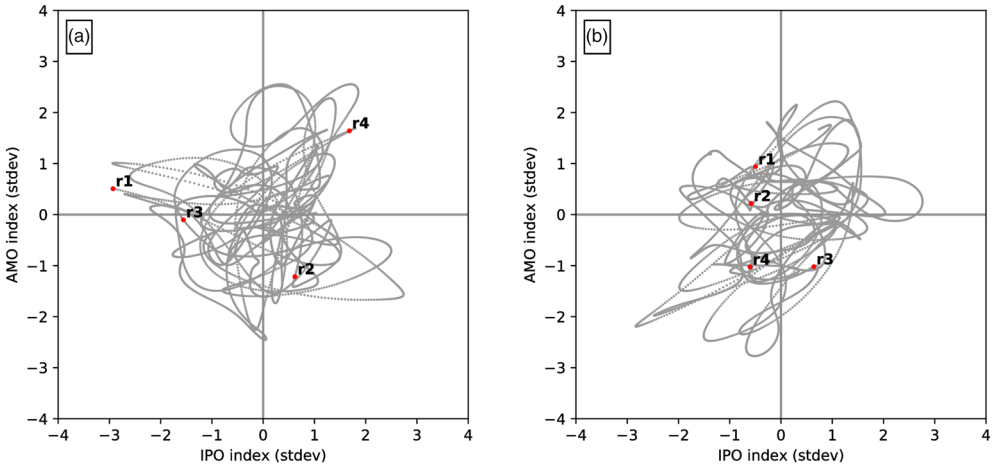

Figure 2. IPO: AMO phase-space diagrams (a) for the GC31-LL *piControl* experiment and (b) for the GC31-MM *piControl* experiment. Labels r1 to r4 indicate the IPO and AMO phase contained within the initial conditions chosen for each member of the historical experiments.

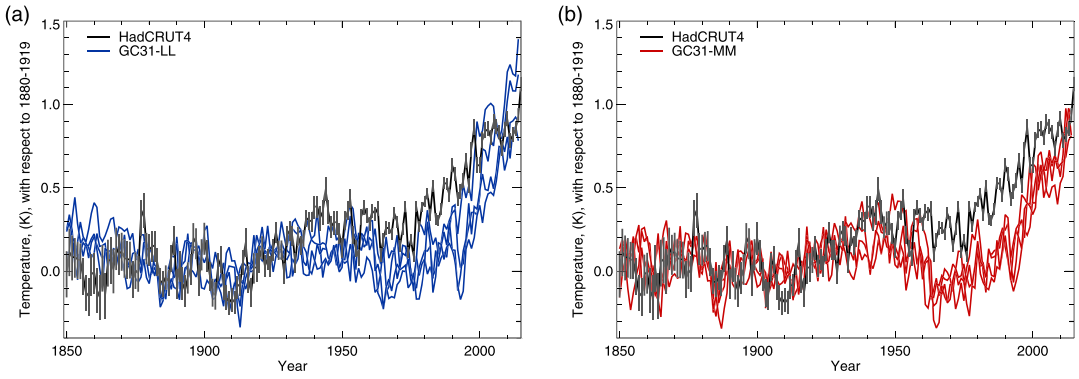

Figure 3. Global mean near surface temperature evolution for (a) GC31-LL (blue) and (b) GC31-MM (red), detrended with estimate of model drift. Uncertainties in HadCRUT4 global annual means are represented by small vertical gray bars. See main text for method.

We also explore the upper OHC of the historical simulations. The heat absorbed and released by the ocean is an important factor when considering the Earth's energy budget. The relatively short ocean subsurface observational record, which covers the near-present-day warming period, provides a benchmark against which the model ocean heat uptake (OHU) can be assessed.

Furthermore, we examine decadal variability of precipitation in our experiments to understand whether notable historical wet and dry epochs can be replicated, which would then suggest a link to historically evolving forcings.

2.1. Global Mean Near-Surface Air Temperature

Global mean surface air temperatures (GMSAT) at 1.5 m from the historical simulations are compared with an observed temperature record in Figure 3. HadCRUT4 is a gridded blended land air and SST data set, covering the period 1850 to the present day (HadCRUT.4.6.0.0 Morice et al., 2012). The GC31-LL and GC31-MM data are processed following Jones et al. (2013), that is, monthly 1.5-m air temperatures are regridded onto the HadCRUT4 spatial grid, anomalies found with respect to the 1961–1990 period removing the seasonal cycle, masked by the HadCRUT4 coverage so that only grid points with observed data are retained, anomalies found again with respect to the 1961–1990, where 50% of data are needed to create the mean for the reference period. Global annual means are then calculated for the model and observed data and the average over the 1880–1919 period removed to find anomalies relative to that period. The uncertainties on the HadCRUT4 global annual means are given with the data set as 95% error bounds for measurement, grid box sampling, and bias uncertainties. Here they are shown relative to the 1961–1990 period. As the model data are sampled the same as the observations, coverage uncertainty is not included in the figure.

It is common for studies to attempt to remove model drift from data in plots of transient changes in model diagnostics, with different approaches having pros and cons (Sen Gupta et al., 2013). Many studies do not try to remove the drift, as what may be assumed to be model drift could be dominated by internal model variability on long times (Jones et al., 2013). Here an estimate of the models' surface air temperature drifts is subtracted from the global mean historical temperatures. The *piControl* simulations (Menary et al., 2018) have long-term trends of 0.026 K/100 years (GC31-LL *piControl*) and 0.110 K/100 years (GC31-MM *piControl*), with full spatial coverage. To account for the changing spatial coverage of HadCRUT4, the following is done. The spatial trend through the first 500 years of the *piControl* experiments (Menary et al., 2018) is deduced. The trend over the 1850–2014 period is reconstructed and then processed as for the historical simulated data above. The global mean is removed from the historical temperatures, thus detrending with an estimate of the models' drifts.

The GC31-LL historic simulations warm 0.79 ± 0.32 K for the 2000–2014 period relative to 1880–1919. GC31-MM warm 0.64±0.07 K over the same period. This compares with the observed warming of 0.83±0.05 K. Uncertainties are ±2 standard deviations.

While the overall trend over the historic period is similar between the observations and the models, both models show less warming than observed over the first half of the twentieth century and somewhat more rapid warming than observed over the last 20 years of the period. One has to be careful in how to interpret

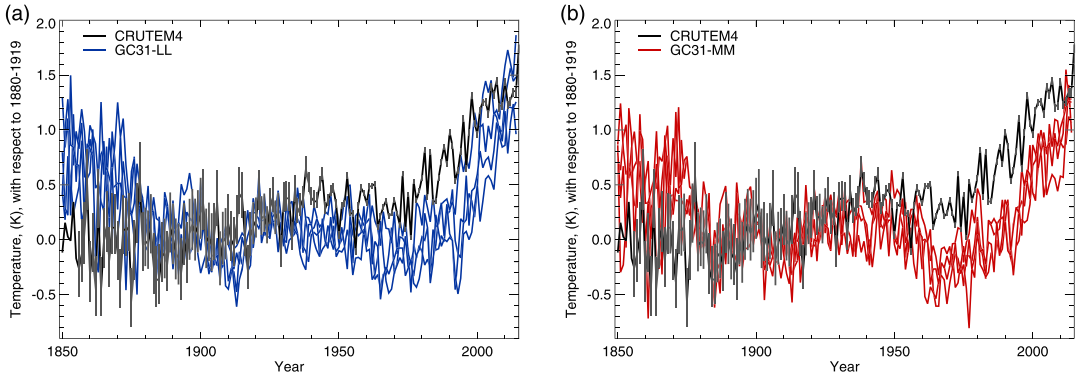

Figure 4. Global mean land temperature evolution for (a) GC31-LL (blue), and (b) GC31-MM (red). Near-surface air temperature, with all data on same grid, masked by model land mask and masked by CRUTEM4. Detrended with estimate of model drift. Uncertainties in CRUTEM4 data are represented by small vertical gray bars. See main text for method.

differences between the simulations and observed temperature record, especially when not examining a sample of models (Jones et al., 2013). The choice of reference period can make a substantial difference in when differences between models and observations are apparent (Hawkins & Sutton, 2009).

Superimposed onto the longer-term trends in GMSAT are multiyear temperature dips associated with the injection of volcanic aerosols into the atmosphere during major tropical volcanic eruptions, in particular 1883 (Krakatau), 1902 (Santa Maria), 1963 (Mount Agung), 1982 (El Chichón), and 1991 (Mount Pinatubo). Analysis of the impact of these eruptions is discussed further in section 5.

2.2. Global Mean Land Near-Surface Air Temperature

The global land surface air temperature (GMSAT-Land) evolution of the two historical experiments is shown in Figure 4. The processing is the same as that for GMSAT in Figure 3, with an additional step of applying a land mask to the data, before the global mean is calculated. The land mask is constructed from GC3-LL's land area fraction, regridded onto the 5° × 5° grid and set to missing data value where land area is less than 25%. The land air temperature data set, CRUTEM4, is used (CRUTEM.4.6.0.0 Jones et al., 2012), and the models are masked by its spatial coverage. For consistency, CRUTEM4 is also masked by the same land mask used by the models. As for HadCRUT4, the uncertainties on the CRUTEM4 global annual means are given with the data set as 95% error bounds for measurement, grid box sampling, and bias uncertainties.

The GMSAT-Land increase, between the 1880–1919 and 2000–2014 periods, is smaller for the models than observed. GC31-LL warms by 1.03 ± 0.48 K and GC31-MM by 0.85 ± 0.20 K, compared to the observed warming of 1.27 ± 0.22 K. Both models, again, show a stronger cooling over the 1950–1970 period, followed by a stronger warming than observed. The models also show a strong cooling between 1850 and 1879. A comparison of the GC31-LL cooling trend over the period 1850–1879 with the equivalent 30-year segments from the *piControl* (30 years following the date of each initial condition), using a difference of means Student's *t* test, indicates that the cooling is marginally significant ($p = 0.04$). CRUTEM4 has limited coverage before 1900, being concentrated over North America and Europe; hence, it is probable that this is a representation of regional changes rather than of a global signal. A larger ensemble size would be necessary to confirm that this is more than an artifact of small ensemble size and poor observational coverage in the early part of the instrumental record.

2.3. Global OHU

About 93% of the current Earth's energy imbalance enters the world ocean as an increase in OHC. This makes OHC one of the most relevant metrics of climate change, arguably more relevant than surface temperature records (Von Schuckmann et al., 2016).

The fifth Coupled Model Intercomparison Project (CMIP5) General Circulation Models displayed a wide range of OHC change in the upper 700 m over the late twentieth century, ranging from 30 to 280 ZJ for the period 1971–2005 (Flato et al., 2013, Figure 9.17), with the observational data sets indicating between 100 and 120 ZJ, where 1 ZJ is 10^{21} J. Differences across the General Circulation Models in the representation

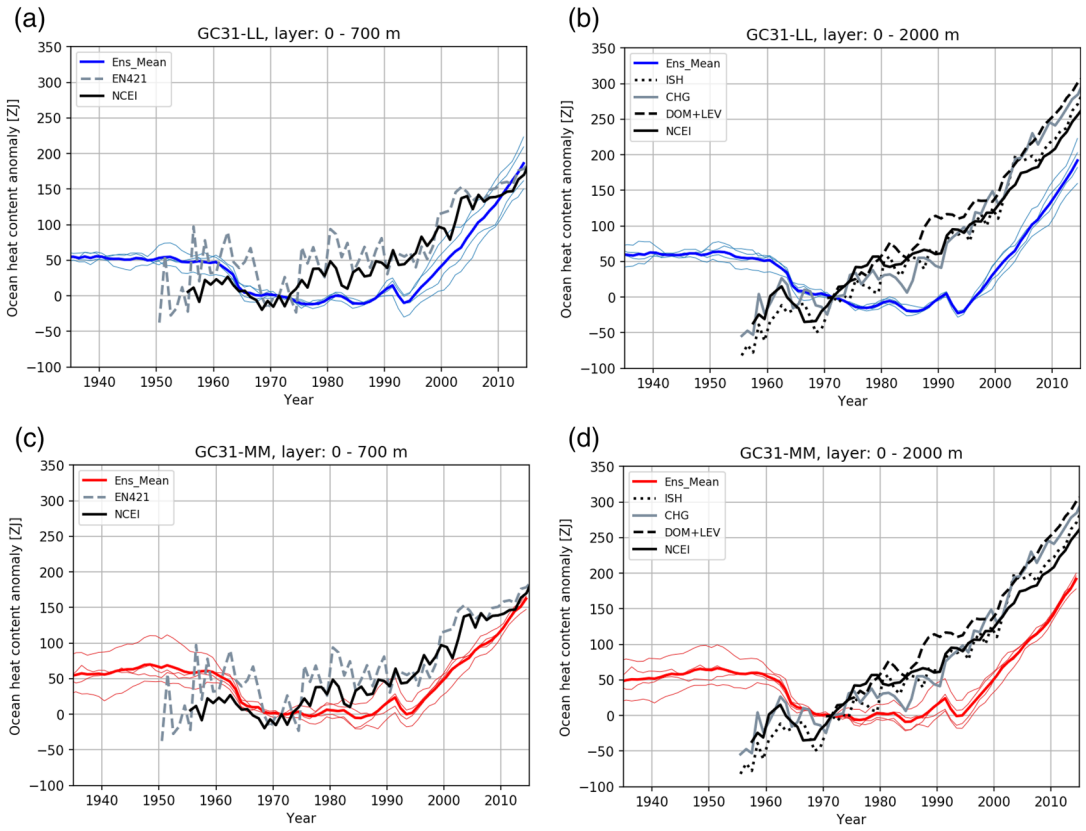

Figure 5. Time series of annual mean global ocean heat content anomaly for GC31-LL (a, b) and GC31-MM (c, d) in the 0- to 700-m layer (a,c) and the 0- to 2,000-m layer (b, d). Ensemble means in bold color (blue for GC31-LL, red for GC31-MM); individual ensemble members in thin colored lines; various observational products in thick gray and black lines. All time series normalized to the year 1971. In the 0- to 2,000-m panels the NCEI data are 5-year running means.

and parameterization of OHU processes explain this wide range. We have used various observational data sets of OHC to evaluate the two historical experiments against. These are the following: EN4.2.1, an updated version of Good et al. (2013); NCEI, an updated version of Levitus et al. (2012); ISH, an updated version of (Ishii et al., 2017); CHG, Cheng et al. (2017); and DOM+LEV, an updated combination of Domingues et al. (2008) and Levitus et al. (2012). We consider time series of global and basin-wise OHC in the upper 700 and upper 2,000 m which are readily available for the observational data sets.

We discuss the OHC drift in the *piControl* simulations (Menary et al., 2018) first. In the global average, both GC31-MM and GC31-LL are warming at all levels in the *piControl*: 0–700, 0–2,000, and the full-depth global ocean volume (not shown). For the full-depth global ocean, the warming rates in the *piControl* simulations appear fairly linear, amounting to about 200 ZJ/100 years in GC31-LL and 400 ZJ/100 years in GC31-MM. Hence, we decided to use linear detrending of the historical simulations. For each simulation and layer depth individually we estimated the linear trend from the matching part of the *piControl* and subtracted it.

For the global OHU in the layer 0–700 m and a reference year of 1971, both sets of the GC31 simulations do not warm in the 1970s and 1980s, as opposed to the observational data sets that show a warming of about 50 ZJ during the 1970s (Figures 5a and 5c). After 1993, the simulations warm at a faster rate than the observations suggest. For the overall time period from 1971 to 2014, both GC31-LL and GC31-MM simulate an OHC change very close to observations, especially given the wide spread among the CMIP5 models (Flato et al., 2013, Figure 9.17). The GC31-LL ensemble mean warms by about 25 ZJ more than the GC31-MM ensemble mean in this time period, which is consistent with a warmer global mean near-surface temperature and a higher near-surface air temperature warming rate for GC31-LL compared to GC31-MM (Figure 3).

In the layer 0–2,000 m (Figures 5b and 5d), the GC31-MM ensemble mean shows zero OHC change between 1971 and 1993, and the GC31-LL ensemble mean even shows a slight heat loss, while the observations suggest a continuous warming of about 100 ZJ in this time. This is in line with the higher warming rate simulated

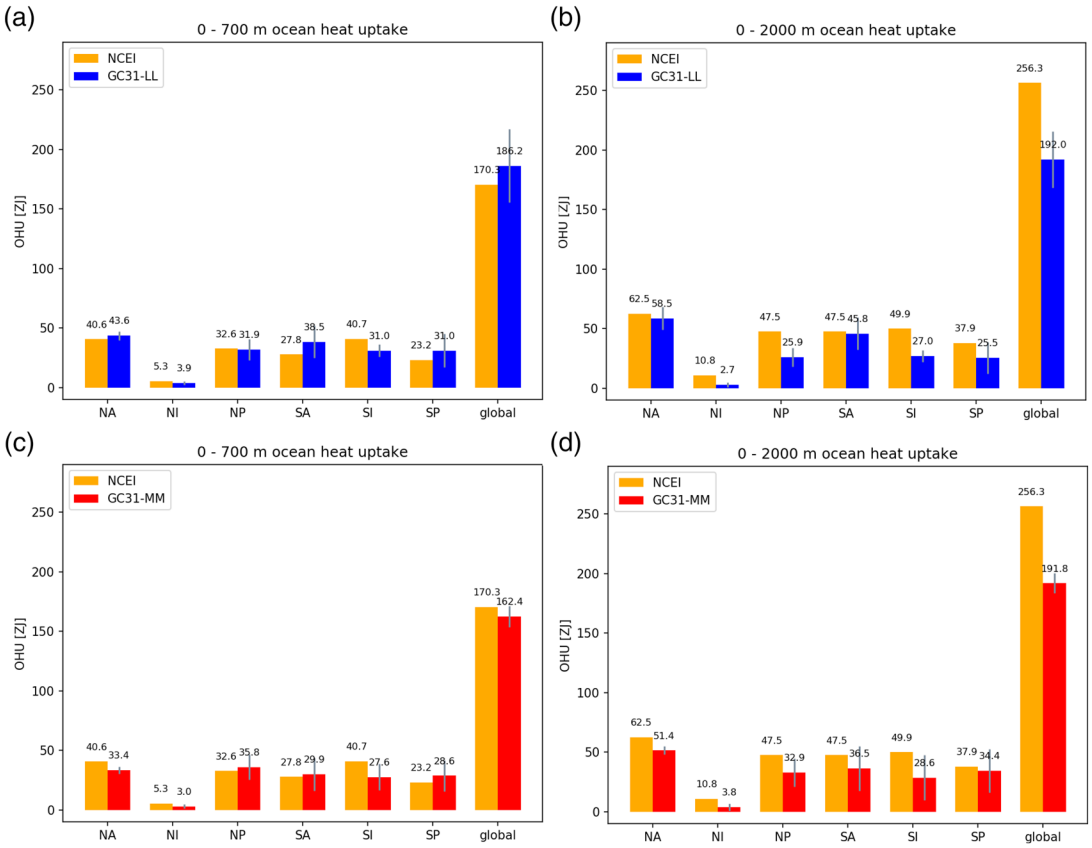

Figure 6. Total ocean heat uptake (change in ocean heat content), 1971–2014, for six individual basins (Northern/Southern Atlantic/Indian/Pacific) and the global ocean for GC31-LL (a, b; blue) and GC31-MM (c, d; red) in the 0- to 700-m layer (a, c) and the 0- to 2,000-m layer (b, d). The narrow gray bars indicate ±1 standard deviation of the ensemble. Orange bars: observational estimate of OHU from the NCEI data set.

in GC31-MM near-surface air temperature in this period. From 1993 to 2014, the simulated warming rates are very close to the observed ones. Since too much heat is taken up in the 0- to 700-m layer in this period, it follows that the warming rate below 700 m must be smaller than observed.

For GC31-LL, we note that the four-member ensemble members are split into two pairs with distinctly different warming rates, especially in the 0- to 700-m layer (Figure 5a). Andrews et al. (2019) noted the accompanying, distinctly different development of Southern Ocean sea ice cover in these two simulation pairs (their Figure 8). We believe that this is related to a temporary increase (blip) in the Southern Ocean deep water formation rate, revealed by an associated positive blip in the strength of the ACC, compare with Figure 12 in Menary et al. (2018), and followed by a more stably stratified Southern Ocean. Two of the GC31-LL simulations branch off the *piControl* before this blip and two after. The simulation pair with the smaller warming rate in Figure 5a is the one branching off after the blip, indicating a more stable stratification and hence less OHU in the Southern Ocean. We also find that the pair with the smaller warming rate in the late twentieth century has a smaller cooling rate in the years 1850 to about 1950, indicating again a more stable ocean stratification and hence less heat loss.

2.4. Regional OHU

The globally integrated trends in OHC discussed above mask distinctly different trends in the individual basins. To analyze these regional trends, we have separated the three major basins (Atlantic, Pacific, and Indian) into a northern and southern parts, cutting at the equator and disregarding marginal seas (e.g., Hudson Bay and the Mediterranean). Figure 6 shows the total change in OHC from 1971 to 2014 in these six individual basins.

In the 0- to 700-m layer (Figures 6a and 6c) the simulated global OHU is rather close to the observational estimate, with GC31-LL taking up about 15% more heat than GC31-MM and having a larger spread as discussed

above. In both models there is a tendency to underestimate the OHU in the South Indian Ocean and to overestimate it in the South Pacific. The larger OHU in GC31-LL happens mostly in the North and South Atlantic basins. This might well correlate with the stronger Atlantic Meridional Overturning Circulation (AMOC) in GC31-LL in the late 20th and early 21st centuries (Figure 17; see also discussion in section 4).

By contrast, both models show almost exactly the same OHU in the 0- to 2,000-m layer globally (Figures 6b and 6d), with the simulated values being much smaller (25%) than in the observations. In both models, the underestimation of the OHU is largest in the South Indian, North Pacific, and North Indian oceans, while in the two Atlantic basins the underestimation is small or even insignificant.

2.5. Decadal Precipitation

Variability in precipitation is one of the most important ways in which climate can impact human societies, to the extent that dry epochs in semiarid regions can even lead to life there becoming unviable. Assessment of the extent to which the historical model ensembles reproduce these variations, however, is limited by the availability of observational data. Before the era of satellite observations (prior to the late 1970s), there were essentially no marine precipitation estimates, so comparisons are limited to land-based data derived from rain gauges. In addition, the availability of even these data becomes confined to fewer regions further back in time (Becker et al., 2013). As such, it is not possible to make comparisons with the model simulations over their full period from 1850. Here we present an analysis for the period from 1910 to 2010. The focus is on decadal anomalies as these highlight eras of persistent wetness or drought which have more chance of being a response to the imposed external forcings than year-to-year precipitation variations, which are likely to have a substantial component of internal variability.

Observed decadal mean precipitation anomalies, derived from the Global Precipitation Climatology Centre (GPCC) monthly data set (Schneider et al., 2011), are shown alongside the GC31-LL and GC31-MM ensemble means for equivalent decades in Figure 7. It is noticeable that observed regional precipitation variability is generally larger and more spatially heterogeneous than in the ensemble means of the historical simulations. The overall level of correspondence between the model means and observations is generally low, although the correspondence between the simulations at the two model resolutions is much greater. This suggests that the two model resolutions respond to a common influence, namely, the external forcings. In contrast, the dissimilarity of the observed decadal means suggests that decadal rainfall anomalies are mainly a result of internal climate variability (Scaife et al., 2009). This is consistent with previous work, which has suggested that decadal or multidecadal modes of variability are important in generating decadal regional precipitation variability (e.g., Folland et al., 1986; Power et al., 1999). Despite the low level of agreement between the models and observations, there are some similarities; the drying of the African Sahel region seen from the 1970s is present to some extent. This suggests at least a contribution from forcings, as in CMIP5 historical model simulations (Biasutti, 2013) and as proposed in response to anthropogenic aerosols (Booth et al., 2012). In the simulations presented here, this feature appears to be part of a wider drying of many parts of the tropics and subtropics between the 1960s and 1990s.

To examine regional decadal rainfall in more detail, we show area-averaged decadal precipitation for large- and regional-scale areas in Figure 8. For the globe as a whole, the Northern Hemisphere (NH), the Tropics, North America, Africa, and Asia in the models (at both resolutions), there is a tendency for the first half of the twentieth century to be wetter than the 10-decade average, with a drier-than-average second half of the century. At the end of the period there is a recovery to near average or wetter conditions. This typically does not correspond well with the temporal changes in many of these regions, which tend to show wetter conditions following a dry early twentieth century. Exceptions include the Tropics and Africa, although the onset of drying occurs earlier in the models. For South America, observations show periods of wetter and drier conditions, but the models show a long-term drying trend, albeit with a high ensemble spread, such that precipitation anomalies in most decades fall within the model range. For Europe, there is very little decadal variability in the model, inconsistent with observations that show climate tending wetter since the 1940s.

At the smaller scale, we examine regions that have a tendency for drought. As already mentioned, Sahel rainfall is suppressed in the models in the 1970s (and less clearly in the 1980s), although to a much lesser degree than in observations. The simulations also show the previous two decades to be dry, however, in contrast to wetter-than-average conditions in observations. A similar relationship is found in West Africa. There is a simulated increase in precipitation in the 2000s in East Africa, whereas the observations show

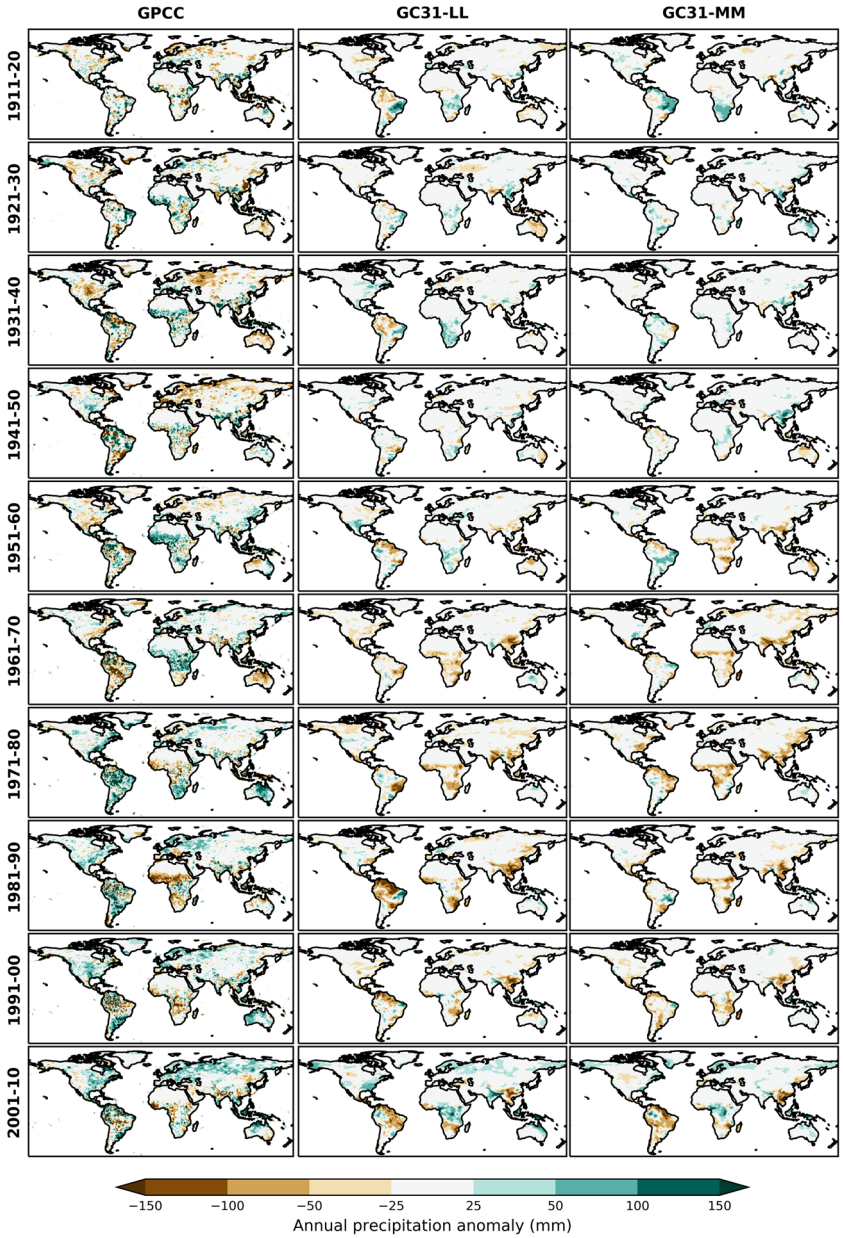

Figure 7. Decadal mean precipitation anomalies (mm/year) for observations (derived from the Global Precipitation Climatology Centre (GPCC) monthly data set of Schneider et al., 2011) and the means of the GC31-LL and GC31-MM ensembles.

continued drying. In the Indian monsoon region, there is also a positive trend in rainfall simulated in recent decades that is not seen in observations. In central China, there is some indication of the observed drying from the 1940s to the 1960s in the models, but the recovery in the 1980s and 1990s is not reproduced. There is no indication of the South East Australian drought of the early 2000s in the model simulations, although observations are not outside the spread of the ensemble for at least one model resolution. The famous Dust Bowl drought of the 1930s on the North American Great Plains is also inconsistent with the forced responses of the models, as is the moist epoch of recent decades. Rainfall in the Brazilian Nordeste region shows large multidecadal variations that have been linked to the Atlantic Multidecadal Variability (AMV; Knight et al., 2006). Although the simulations capture part of the observed AMV (see section 4), in the model this does not appear to act as a significant control on rainfall in these regions, as is implied by observations. Decadal rainfall over the Amazon does not appear to reflect the occurrence of unprecedented drought in specific

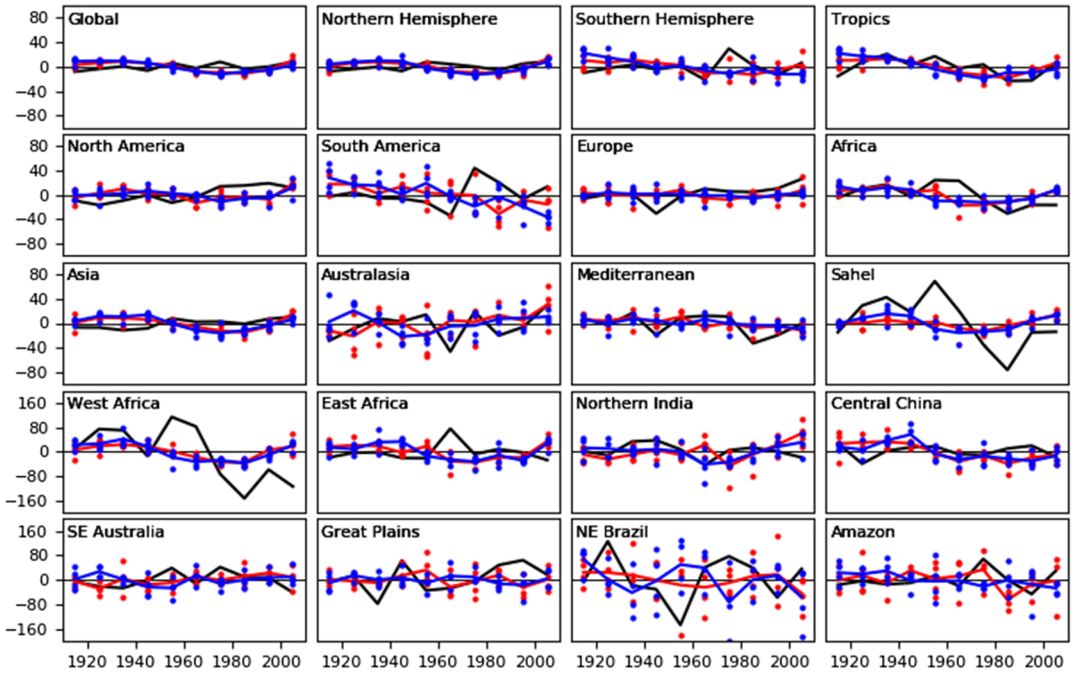

Figure 8. Time series of decadal precipitation anomalies (mm/year) for selected regions. Observed precipitation from the GPCC data set is shown as the black line, while the GC31-LL and GC31-MM experiments are represented by blue and red, respectively. For the models, the solid line is the ensemble mean and the dots are values in individual ensemble members. The regions shown are as follows: Global (180°W to 180°E, 60°S to 90°N); Northern Hemisphere (180°W to 180°E, 0–90°N); Southern Hemisphere (180°W to 180°E, 60–0°S); Tropics (180°W to 180°E, 23°S to 23°N); North America (180–60°W, 25–90°N); South America (85–30°W, 60°S to 15°N); Europe (35°W to 50°E, 35–75°N); Africa (20°W to 55°E, 40°S to 40°N); Asia (50–180°E, 5–90°N); Australasia (105–180°E, 50°S to 10°N), Mediterranean (15°W to 40°E, 30–45°N); Sahel (15°W to 37.5°E, 12.5°S to 17.5°N); West Africa (20°W to 10°E, 0–10°N); East Africa (30–50°E, 10°S to 10°N); Northern India (70–90°E, 17.5–30°N); Central China (90–120°E, 25–35°N); South East Australia (137.5–155°E, 40–30°S); North American Great Plains (105–95°W, 33–40°N); North East Brazil (50–35°W, 10–0°S); Amazon (75–50°W, 10°S to 5°N). Note that spatial averages are computed with area-weighting and marine areas masked out.

years in the 2000s, but the simulations do contain the forced decline in rainfall that is seen for the South American continent as a whole.

3. Sea Ice, Ice Sheets, Permafrost, and Snow Cover

Aspects of the cryosphere are sensitive to changes in global and local temperature variations on a range of timescales. For example, over recent years there is evidence of a reduction in Arctic sea ice extent, a decline in the mass of the Greenland ice sheet, a reduction in snow cover, and a decline in permafrost extent. This has a potential impact on the Earth's climate system by influencing the surface energy budget, sea levels, and atmospheric and oceanic circulations. See Vaughan et al. (2013) for a comprehensive assessment of trends relating to the cryosphere and the impacts on the climate system.

Here we consider a number of components of the cryosphere in the historical simulations, which include sea ice extent and volume, Antarctic and Greenland ice sheet surface mass balances, permafrost, and snow cover extent. We investigate whether the simulated response to climate variability is similar to observations.

3.1. Near-Present-Day Sea Ice, 1990–2009

Figures 9 and 10 compare 1990–2009 average summer and winter sea ice concentration and extent, in both hemispheres, for GC31-LL and GC31-MM, respectively with the HadISST.2.2 data set of Titchner and Rayner (2014). The modeled and observed sea ice extent—the total area of ocean with at least 15% ice cover—are included as annotation.

In the Arctic there is generally good agreement between the modeled (shading) and the observed (orange lines) location of the ice edge as depicted by the 15% concentration contour. In the summer the Arctic sea

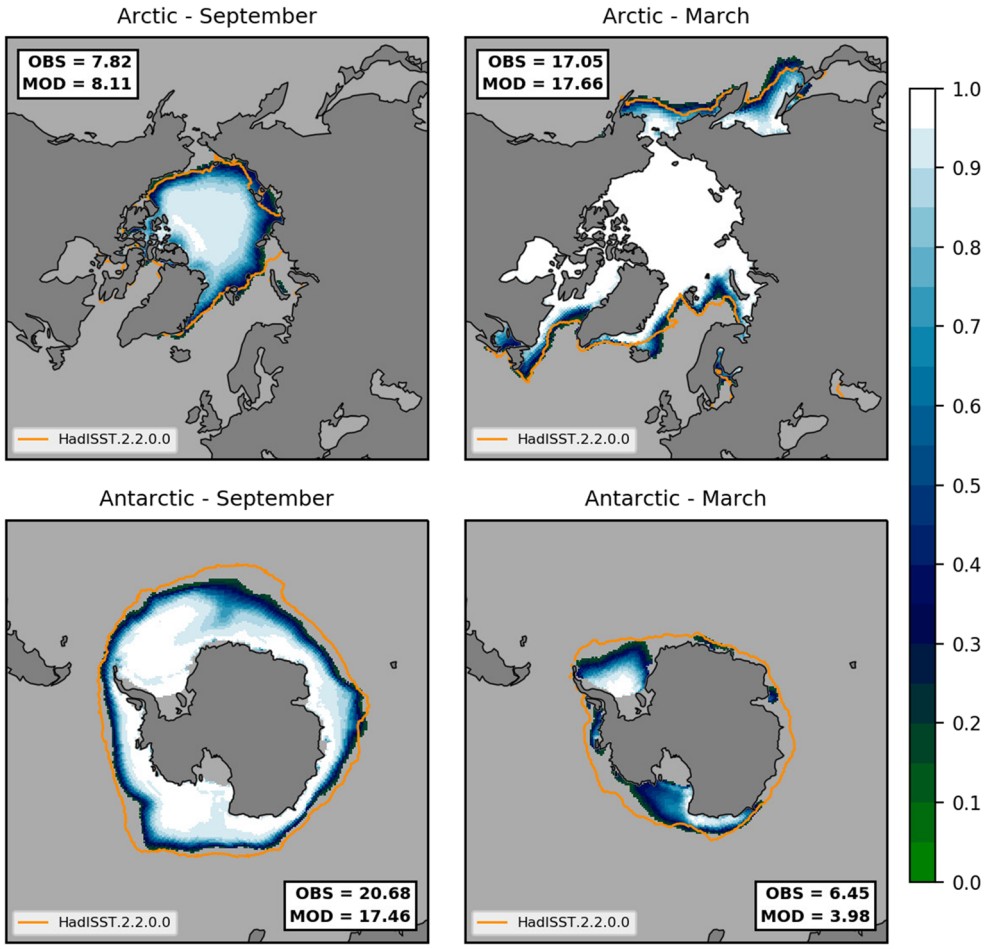

Figure 9. GC31-LL historical ensemble mean 1990–2009 Arctic (top row) and Antarctic (bottom row) sea ice concentration (shading) for September (left column) and March (right column). Areas with sea ice less than 15% are masked out. Also shown (orange lines) are corresponding 15% sea ice concentration contours from the HadISST.2.2.0.0 data set (Titchner & Rayner, 2014). Inlaid boxes show the observed and modeled sea ice extent (units: 10^6 km^2).

ice is in very good agreement with HadISST.2.2 within the Arctic Ocean basin, except for in the eastern Kara Sea where the sea ice cover in both resolution models is too low. The winter ice extent in the GC31-LL configuration is also in good agreement with HadISST.2.2, but there are a few differences in the GC31-MM configuration, which has too little winter ice in the Sea of Okhotsk and the Labrador Sea and a larger overestimation of ice north of Iceland. The GC31-MM model has a more realistic path for the North Atlantic current (Menary et al., 2018) and the deep convection occurs principally in the Labrador Sea, while in GC31-LL the deep convection occurs south of Svalbard (Kuhlbrodt et al., 2018). Consequently, the GC31-MM winter sea ice is less extensive in the Labrador Sea and more extensive in the Greenland Sea.

In the Antarctic the GC31-MM model sea ice is under extensive, which is caused by the Southern Ocean warm bias being much higher in the ORCA025 ocean used in the GC31-MM configuration (Hyder et al., 2018). The GC31-LL model has a weaker Southern Ocean warm bias leading to more ice and hence better agreement with HadISST.2.2.

Although much more realistic than for the higher-resolution GC31-MM configuration, Southern Hemisphere sea ice in the GC31-LL configuration is still a little less extensive than observed. In summer ice is low in the Amundsen Sea, the north and eastern Weddell Sea, and most of east Antarctica; in winter ice cover is low in the eastern Ross Sea and east of the Weddell Sea sector south of Africa—the regions of strongest Southern Ocean warm bias (Kuhlbrodt et al., 2018).

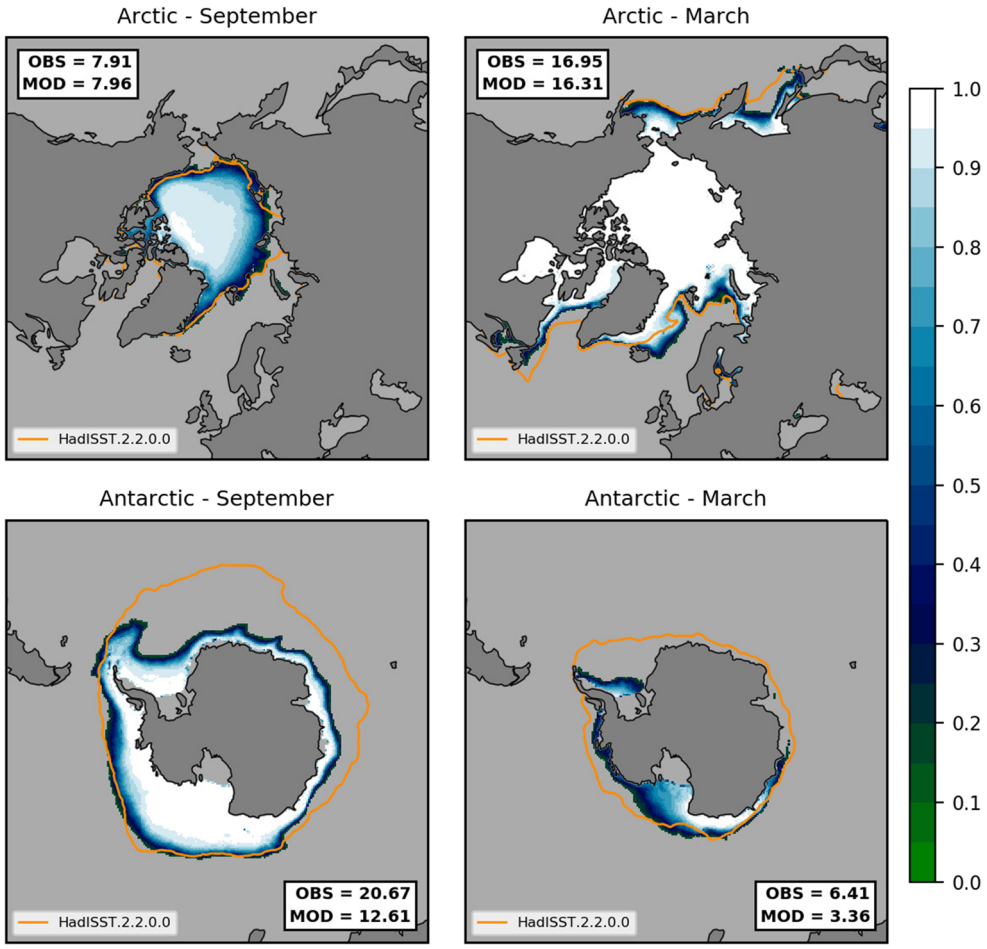

Figure 10. As in Figure 9 but for GC31-MM. Note that the observed extent values are different than those reported in Figure 9 because they are calculated independently for each resolution using the relevant model grid.

3.2. Evolution of Sea Ice Extent and Volume, 1850–2014

Figure 11 shows the evolution of summer and winter sea ice extent in both hemispheres. Total basin-wide NH sea ice extent (Figure 11a) from GC31-LL agrees well with HadISST.2.2 in both March and September. Winter NH extent in GC31-MM is lower than for GC31-LL and is low compared with HadISST.2.2. This is consistent with Figures 9 and 10, which show that GC31-MM has too little winter ice in the Sea of Okhotsk and the Labrador Sea. This offset is apparent throughout the whole 1850–2014 historical period, in line with the offset in the *piControl*s. However, GC31-MM March sea ice is close to the NSIDC sea ice index (Fetterer et al., 2017) and so these are within observational uncertainty. The September ice extent is very similar for both model resolutions. NH sea ice extent shows some evidence of multidecadal response to the forcing throughout the historical period but is mostly within 1 standard deviation of the 500-year mean of the *piControl*, until around 1995 when the ice starts to decline.

Trend lines and gradients are also included on Figure 11. The 1985–2014 trends in GC31-LL are always higher than in GC31-MM. The summer decline is much larger than the winter decline in both models, consistent with observations. However, the difference between winter and summer trends is not as high as observed. GC31-MM agrees well with observations in the winter while GC31-LL is too large; meanwhile in summer GC31-LL agrees well with observations while GC31-MM is too low.

The Southern Hemisphere sea ice extent (Figure 11b) is roughly constant, within 1 standard deviation of the *piControl*, throughout most of the historical period until around 1985 when the winter ice starts to decline. This is contrary to the observational data sets, which show a small increase in Antarctic extent for this period. The extent increase associated with the mean observed trends over the 1985–2014 period is generally smaller

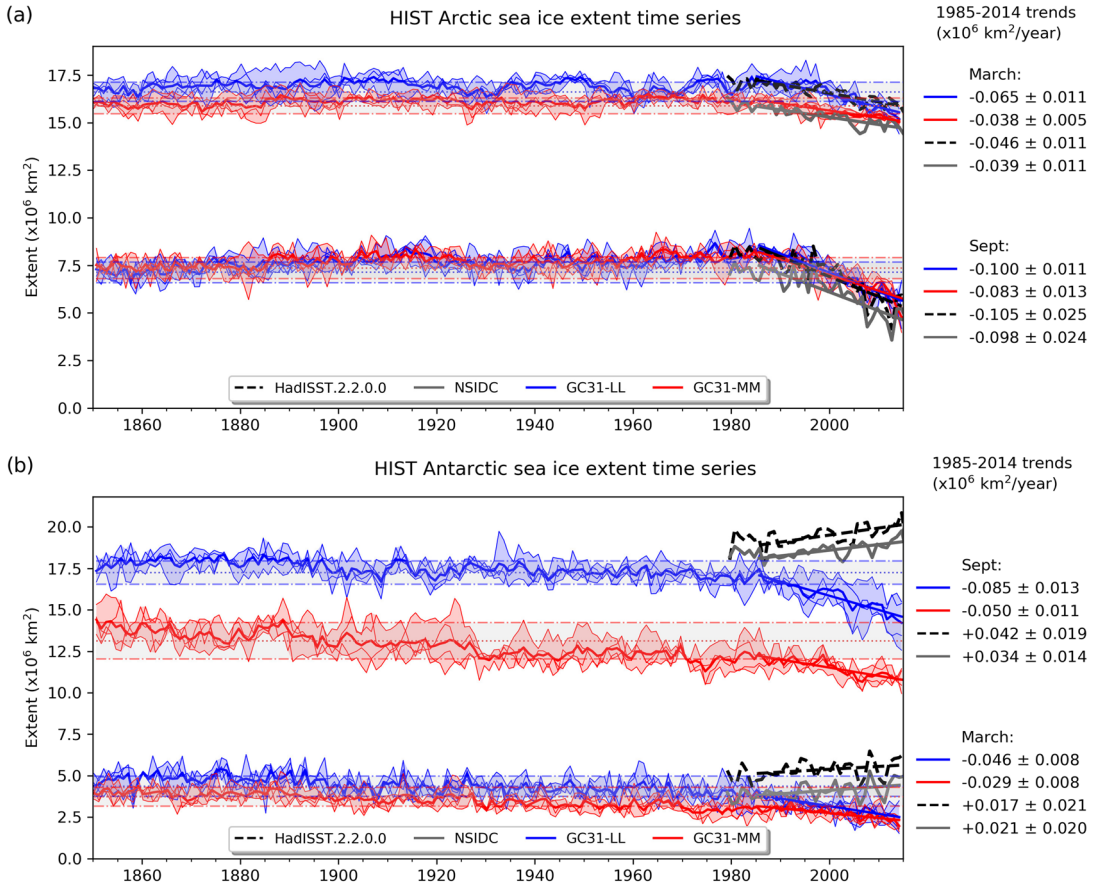

Figure 11. Time series of basin-wide sea ice extent from the GC31-LL (blue) and GC31-MM (red) models throughout the 1850–2014 historical simulations compared with HadISST.2.2.0.0 and NSIDC. (a) Northern Hemisphere March (upper) and September (lower). (b) Southern Hemisphere September (upper) and March (lower). In all panels individual members are plotted with thin colored lines and ensemble means are the thick colored lines. Colored shading is added between the ensemble minimum and maximum. The long-term mean from the corresponding *piControl* is plotted as the colored dotted line; ±1 standard deviation is indicated by the colored chain lines with gray shading in between. Linear regressions for the ensemble mean 30-year period 1985–2014 are included; trend slopes with 95% confidence intervals are displayed to the right of each panel. (unit: 10^6 km^2).

than 1 standard deviation of the *piControl* simulations, except in winter (September) when the HadISST.2.2 increase is above 1 standard deviation of the GC31-LL *piControl*. However, the extent increase associated with the 95% confidence interval of the observed trends is quite large and spans the *piControl* standard deviation.

Arctic sea ice volume is an important climate indicator because it can be considered a consequence of the integrated surface energy balance of the Arctic Ocean. As sea ice volume depends on both ice thickness and areal extent, it is therefore more directly tied to climate forcing than the areal extent alone. Figure 12 shows annual mean NH sea ice volume for both model configurations compared with the PIOMAS reanalysis of Zhang and Rothrock (2003). NH sea ice volume increases through the historical period, which is a consequence of the aerosol-induced cooling of the planet common to both GC31 configurations (Bodas-Salcedo et al., 2019; Mulcahy et al., 2018). With extent remaining fairly level, this means that thickness is increasing until around 1985 when it starts to decline rapidly. Over the 30-year period from 1985–2014, the volume trend is larger than for the PIOMAS reanalysis. There is much larger internal variability in sea ice volume than extent, attributable to the fact that Arctic winter ice extent is constrained by the coastline and SSTs. Year-to-year volume is well correlated with hemispheric and global atmospheric temperature (not shown).

3.3. Antarctic and Greenland Ice Sheets

Ice sheets gain mass through snowfall and deposition and lose mass through sublimation, surface melt, and subsequent meltwater runoff. This component is known as the surface mass balance (SMB) and is positive when the ice sheet is gaining mass (Lenaerts et al., 2019). The Greenland and Antarctic ice sheets in

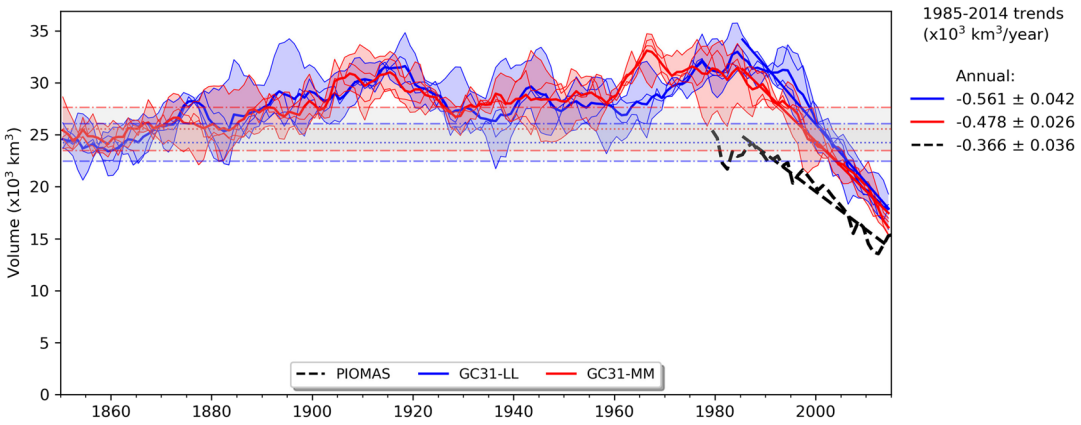

Figure 12. As in Figure 11 but for Northern Hemisphere annual mean sea ice volume compared with the PIOMAS reanalysis (Zhang & Rothrock, 2003). (unit: 10^3 km^3).

HadGEM3-GC3.1 are static, that is, the model does not include ice sheet dynamics. The surface snow is depicted by a multilayer-layer scheme with prognostic grain size and associated albedo (Walters et al., 2019). The snow is initialized to several hundred meters depth to permit continued melt in all future scenarios. Lacking an underlying ice component, if the snow were to be depleted, then no further runoff from melt would be generated. The ice sheet loss, mechanisms of ice shelf basal melt, and iceberg calving are depicted as a freshwater flux at the ice shelf front and a source of Lagrangian icebergs at glacier and ice shelf fronts. The spatial distribution of ice shelf basal melt and iceberg calving is based on climatological observations. The overall fluxes are then weighted to balance the SMB of the *piControl*. The weighted fluxes are held constant throughout the historical and future scenarios as there is no model mechanism to predict changes. Through this mechanism, the global ocean maintains mass and salinity throughout the control. If the ice sheet SMB changes in historical and future scenarios, then the ocean mass and salinity respond accordingly. In Antarctica there is little runoff because the environment is extremely cold (Lenaerts et al., 2012), and as a consequence the SMB is positive. An estimation of the Antarctica SMB, through regional models forced by ERA-Interim, is approximately 2,200 Gt/year (Agosta et al., 2019). It is anticipated that increasing regional temperatures will increase snowfall and the SMB will consequently increase. In Greenland, runoff is an important process of mass removal (Ettema et al., 2009). Between 1961 and 1990, the SMB is calculated to be approximately 450 Gt/year (Ettema et al., 2009). Rising temperatures are expected to increase melting faster than increasing snowfall, and the Greenland SMB is expected to decline, likely becoming negative this century.

The SMB calculated from the historical simulations is shown in Figure 13 for both the Greenland and Antarctic ice sheets and for both model resolutions. Variations in the SMB are indicators of global temperature change which influences both snowfall and melt. In observations, Greenland shows a sharp decline in SMB from an average of 422 ± 10 Gt/year prior to 1990 and an average of 286 ± 20 Gt/year from 2000–2018, indicating increased melting (Mouginot et al., 2019). Over the same historical periods, GC31-MM has 431 and 302 Gt/year (2000–2014) and GC31-LL has 307 and 191 Gt/year, respectively. The Antarctic SMB remains constant through the historical period with observations indicating a value of 2,200 ± 111 Gt/year (Agosta et al., 2019). Over this period, the SMB of GC31-LL is 1,842 Gt/year and GC31-MM is 2,049 Gt/year. Thus, GC31-MM agrees well for both ice sheets, within the observational uncertainty, but GC31-LL is producing a rather low SMB. This pattern is consistent with increased orographic snowfall over the steeper coastal slopes at higher resolution (Ettema et al., 2009).

3.4. Permafrost and Snow Cover

The presence or absence of permafrost is, to first order, determined by the annual mean air temperature. Chadburn et al. (2017) constructed a broad relationship between the probability of permafrost occurrence and the annual mean air temperature. The relationship showed that, on average, for any grid cell with a mean annual air temperature less than −4.3°C there is greater than 50% probability of finding permafrost in an equilibrium climate. The WFDEI 2-m air temperature data set (Weedon et al., 2014) has an area of around 15.2×10^6 km^2 (1995–2014) which tallies fairly well with the observed extent of permafrost (Chadburn et al.,

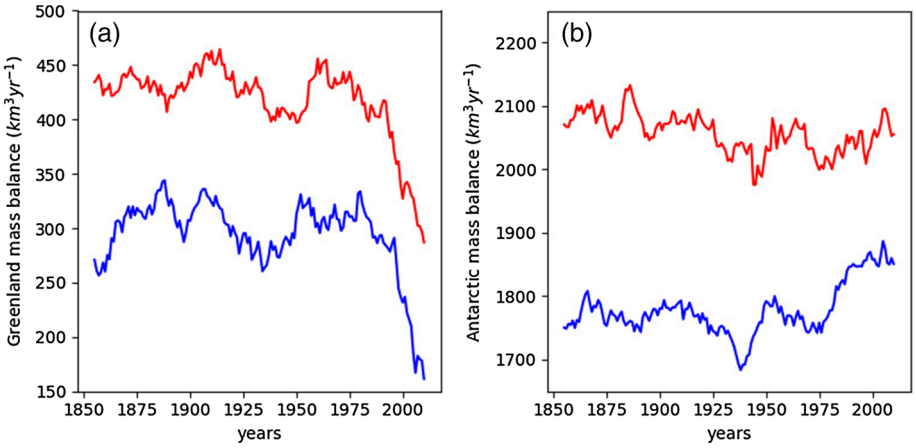

Figure 13. The surface mass balance of the (a) Greenland and (b) Antarctic ice sheets for GC31-LL (blue) and GC31-MM (red). The four-member ensemble means with a 9-year running mean smoothing are shown.

2017). GC31-LL has an area of 16.6–17.9 × 10^6 km^2 and GC31-MM generally has a slightly smaller area of 16.4–16.7 × 10^6 km^2 across the ensemble members over the same period. This suggests that the GC31-MM historical experiment is slightly warmer in the Arctic than the GC31-LL historical experiment. However, both experiments have a colder Arctic than the observations.

Figure 14 shows the change in area of the land surface where the mean annual air temperature is less than −4.3°C. In the regions where mean annual air temperature is less than −4.3°C, it is more than 50% probable that permafrost will exist (Chadburn et al., 2017). The change is not significant in the control runs but is significant in the transient runs and in the observations. The observations show a decrease in area of 0.55–0.58 × 10^6 km^2 per decade. In comparison, the decrease in the model is approximately double, with a loss of 0.9–1.7 × 10^6 km^2 per decade. The GC31-MM experiment warms slower in the Arctic than the GC31-LL experiment, with a slightly smaller reduction in land area with a temperature less than −4.3°C. These results are derived using air temperature alone, but the findings of Chadburn et al. (2017) suggest that they are a suitable proxy for the presence of large-scale permafrost. Therefore, on the basis of air temperature alone, the historical experiments may well predict a greater loss of permafrost than might be expected over the twentieth century. It should be noted that variable land surface characteristics will moderate the deep soil temperatures and hence the local presence or absence of permafrost. This is particularly relevant in the regions with discontinuous or sporadic permafrost. For example, deeper snow will insulate the soil from the cold air in the winter and reduce the probability of permafrost occurrence. Other influencing factors include topography, hydrology, soil properties, and vegetation. Many of these factors are included in the land surface component.

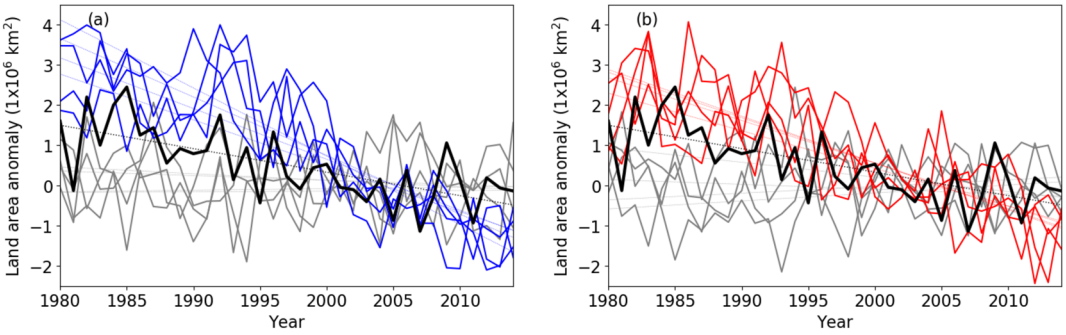

Figure 14. Change in the area of the land surface where the mean annual air temperature is less than −4.3°C for (a) GC31-LL (blue) and (b) GC31-MM (red) ensemble members. The equivalent sections of the *piControl* run are shown in gray and observations (WFDEI) in black. Anomalies are calculated with respect to the period 1995–2014 or their equivalent in the control run.

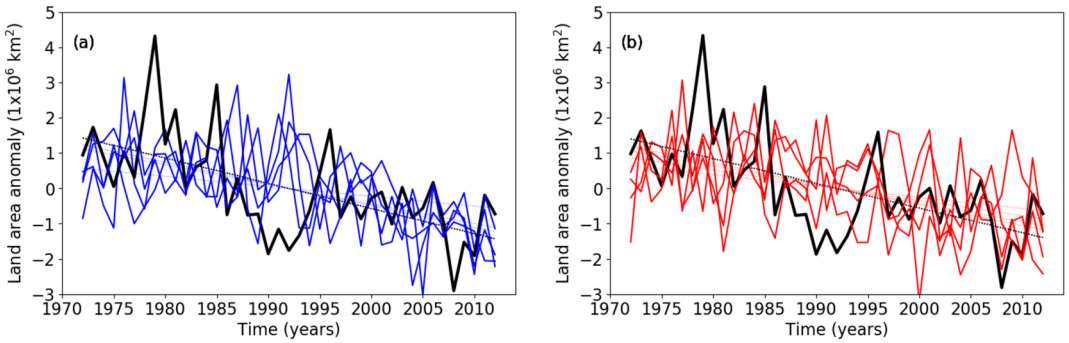

Figure 15. Change in the mean March-April-May snow covered extent for (a) GC31-LL (blue) and (b) GC31-MM (red). Observations (in black) are from Robinson and Mote (2014), and the anomaly is calculated relative to the period 1972–2012.

There has been an observed significant reduction in northern high latitudes snow cover extent of 0.7×10^6 km^2 per decade in March-April-May (Estilow et al., 2015) during the period between 1972 and 2012. The modeled loss of snow cover extent (Figure 15) is slightly lower than the observed loss (Robinson & Mote, 2014) with reductions ranging from 0.28 to 0.71×10^6 km^2 per decade.

4. AMV

We now turn our attention to the North Atlantic in order to consider the representation of the observed AMV. The observed North Atlantic SST has generally evolved differently to the rest of the globe (Sutton et al., 2018). Specifically, it has gone through periods where it warmed faster than the global mean SST (e.g., the 1920–1950s and 1990s–2010s) and periods where it cooled (e.g., the 1960s–1980s). Furthermore, AMV has been linked to many important climate impacts, including rainfall anomalies over Northern Europe and the United States, the strength of monsoons worldwide, and changes in the number of hurricanes (Monerie et al., 2019; Sutton & Dong, 2012; Sutton & Hodson, 2005; Zhang & Delworth, 2006). Most models, including HadGEM3-GC3.1, simulate AMV-like modes of variability spontaneously in their control simulations (Menary et al., 2018; Wills et al., 2019). However, there is evidence that radiative forcing changes have played an important role in shaping the AMV over the twentieth century (Booth et al., 2012; Chang et al., 2010; Watanabe & Tatebe, 2019). Therefore, here we explore evidence for a forced response of AMV in the GC31-LL and GC31-MM historical simulations.

A simple measure of AMV is the basin mean SST anomalies, which we compute over the domain 7.5:75°W and 0:60°N. As simple North Atlantic basin mean AMV indices do not discriminate between local and globally forced signals (Qasmi et al., 2017; Trenberth & Shea, 2006), we also compute an AMV index after removing a global mean signal (hereafter called AMV-GMT). We compute a near-global mean ensemble mean SST index (GMT) by taking the area-weighted mean of all grid points outside the AMV box (7.5:75°W, 0:60°N) and between 60°S and 60°N (hence excluding the poles to permit a more robust comparison with observations). We then remove this ensemble mean GMT index from the ensemble mean AMV index, by linear regression, to produce an ensemble mean AMV-GMT index. We compute the same indices in the historical simulations and in observations (HadISST.1.1 Rayner et al., 2003). The comparison of both AMV indices allows us, therefore, to quantify how much of the variability in the North Atlantic is related to global or local changes.

Using simple basin mean AMV indices, we find that the GC31-LL and GC31-MM historical simulations appear to simulate much of the observed AMV index. Figures 16a and 16b show that both resolutions simulate pronounced multidecadal SST variability within the North Atlantic. The forced basin mean AMV signal, represented by the ensemble mean, is large compared to the ensemble spread. Furthermore, the evolution and timing of the simulated basin mean AMV index is also similar to HadISST, with the correlation of model and observed indices of $r = 0.7$ for both resolutions. However, the basin mean AMV index is not reproduced perfectly. For example, the magnitude of the multidecadal variability is smaller than in HadISST, particularly during the 1940–1960s, and the 1930s warming also appears to occur earlier (by approximately a decade) when compared to observations. Furthermore, the observations lay outside the ensemble spread during the

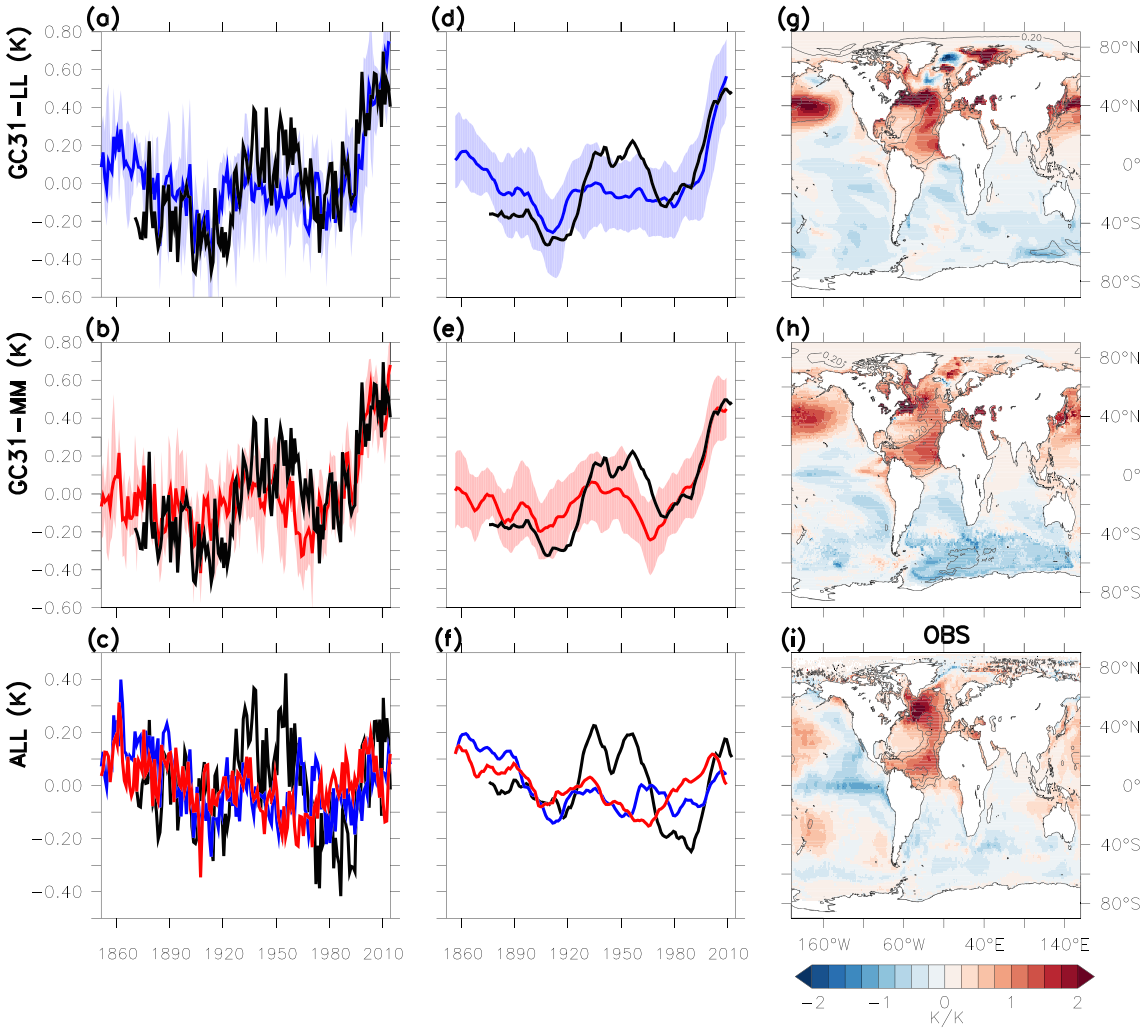

Figure 16. Time series and spatial patterns of AMV in model and observations. (a) Basin mean AMV (SST averaged over 7.5:75°W, 0:60°N) in GC31-LL using annual mean data (blue line: ensemble mean, shading: max/min across ensemble) compared with Observations (black: HadISST Rayner et al., 2003). All indices are anomalies with respect to their 1870–2014 mean (i.e., the observational period). (b) As in (a) but for GC31-MM (red). (c) AMV-GMT: ensemble mean AMV after the global (excluding Atlantic) mean temperature signal (GMT) was removed by linear regression (see text for details) (GC31-LL: blue, GC31-MM: red, obs: black). (d–f) As in (a)–(c) but indices have been smoothed with a 10-year running mean. (g) GC31-LL Ensemble mean SST-GMT regressed onto Ensemble mean AMV-GMT index. SST-GMT: SST after removing GMT by regression (see text). Shading is the coefficient of regression, contours the fraction of the variance explained (square of the correlation coefficient, contour interval: 0.2). (h) As in (g) but for GC31-MM. (i) As in (g) but for observed SSTs.

cooling in the 1920s, and the warming between 1930 and 1950. These discrepancies could show model deficiencies in simulating AMV (i.e., an underestimation of the forced response or internal variability or both). But we cannot rule out that these decades represent an exceptional phase of variability not captured by the small ensemble.

Once the globally forced signal is removed, however, the agreement between simulated and observed AMV (i.e., the AMV-GMT) is much weaker (see Figure 16c). Removing the globally forced signal from the models reduces the magnitude of the forced signal of AMV. In other words, a large component of the total forced AMV signal in the simulations is not unique to the Atlantic. However, this is clearly not the case in observations, where the timing of phases of AMV (i.e., the warm period between 1940 and 1960s and the cool period from 1970s–1990s) is largely unchanged in the AMV-GMT. As a result, correlations between the model and observed AMV-GMT indices drop to $r = 0.2$.

When we consider the spatial fingerprint of the Atlantic local forced response of AMV (i.e., after the globally forced signal is removed), there is evidence for a forced response in the tropical North Atlantic, consistent with previous studies (i.e., Booth et al., 2012; Chang et al., 2010). Figures 16g and 16h show the grid point

linear regression of ensemble mean SST-GMT (SST with the global mean signal (GMT) removed at each gridpoint via linear regression) on the ensemble mean AMV-GMT (Figure 16c) in both models. Both model resolutions show a clear forced signal in the tropical North Atlantic (i.e., with large regression coefficients), and the ensemble mean AMV-GMT index explains more than 40% of the local variance in ensemble mean SST-GMT. In terms of spatial pattern, the tropical fingerprint compares well with that observed (see Figure 16i—here the observed AMV-GMT (Figure 16c) is removed from observed SST via linear regression). However, in the extratropical region, the spatial pattern does not compare well with observations. In the models, the main anomalies are focussed in the Gulf Stream extension region, with little or no signal in the subpolar North Atlantic (between 50°N and 65°N). However, the subpolar North Atlantic is the dominant region of the observed AMV. The extratropics also explains a much smaller fraction of the ensemble mean variance (i.e., < 10%). Given that the internal AMV fingerprint in the *piControl* simulations of HadGEM3-GC3.1 is extratropical (i.e., Menary et al., 2018) in both resolutions, these results support the view that tropical AMV has a locally forced component but that internal variability is more important for the extratropical AMV (Watanabe & Tatebe, 2019).

There is also evidence for a forced response in another important metric of the North Atlantic, the AMOC. Figures 17a and 17d show that the AMOC at 40°N increases from 1850–1990 in both resolutions (GC31-LL: 0.23 Sv per decade, GC31-MM: 0.16 Sv per decade). The simulations also suggest that there is a multidecadal forced AMOC signal that goes beyond the longer timescale forced AMOC trends found in earlier studies (Delworth & Dixon, 2006; Menary et al., 2013). This increase is large compared to the internal variability and moves the AMOC at 40°N outside the range of variability seen in the *piControl*. Following 1990, both model resolutions show a decline in the AMOC. The decline in the AMOC is larger in GC31-MM than in GC31-LL (−1.8 and −0.85 Sv per decade, respectively). Nevertheless, the evolution of the ensemble mean AMOC on decadal timescales is similar between resolutions (correlation $r = 0.7$ after linear detrending and smoothing with a 5-year running mean). AMOC changes are difficult to verify due to a lack of observations, but Smeed et al. (2013) report a −5-Sv per decade trend (5–95% confidence interval (CI): −9.9 to −0.8 Sv per decade) in the AMOC between 2004 and 2012 in the RAPID observations or equivalently a change of −2.7 Sv between 2004–2008 and 2008–2012 (−5.4 Sv per decade). The experiments show smaller ensemble mean trends of −1.4 Sv per decade for GC31-LL (CI: −2.5 to −0.33 Sv per decade) and −0.11 Sv per decade for GC31-MM (CI: −1.8 to 1.6 Sv per decade) over the same period (2004–2012). This is consistent with recent studies (Roberts et al., 2014; Yan et al., 2018) that suggest that the AMOC variability in models may be smaller than is observed. However, the overlapping confidence intervals of the models and observations present some uncertainty in this conclusion.

AMOC changes are usually associated with changes in upper ocean temperatures, and Figures 17b and 17e also show that there is a forced response in the subpolar gyre (SPG) in the North Atlantic (40–60°N) 0- to 700-m average temperatures. At low resolution (GC31-LL, Figure 17b), the upper ocean temperature follows the pattern of cooling and warming seen in the wider AMV-GMT (Fig 16b)—a cooling until the 1970s of about 0.4 K, followed by a rapid warming of about 0.5 K. Interestingly, this evolution compares relatively well with the changes in the observations post-1950, but the ensemble mean does not capture the recent cooling since 2005, which has been in part attributed to a weakening AMOC (Robson et al., 2016). The high-resolution experiment (GC31-MM, Figure 17e) appears to undergo a long-term warming trend modulated by periods of cooling, including after 2000. GC31 does not capture the observed decadal salinity changes in the SPG (not shown); both resolutions show clear long-term trends not seen in observations and do not capture the post-1970s Great Salinity Anomalies. The relationship between upper ocean SPG temperature and salinity will be part of an extended analysis of the SPG for a future study. It is important to note that the upper ocean temperature changes are not only reflecting changes in AMOC (compare Figures 17a and 17b, for example); changes in surface fluxes may also be important. The apparent lag between AMOC changes and changes in SPG mean temperature suggests that the SPG temperatures are driven by a rapid response to surface shortwave fluxes, combined with a lagged response to ocean heat transport.

The forced response in the North Atlantic in the historical experiments (at both resolutions) is largely consistent with previous studies examining the competition between anthropogenic aerosols and the long-term greenhouse gas forcing. For example, Figures 17c and 17f show that the net shortwave absorbed over the subpolar North Atlantic decreased substantially between the 1850s and 1980s, before subsequently increasing. These changes in net shortwave are consistent with sulfate aerosol precursor emissions, which increased over North American and Europe up until the 1980s, before a large subsequent reduction (Booth et al.,

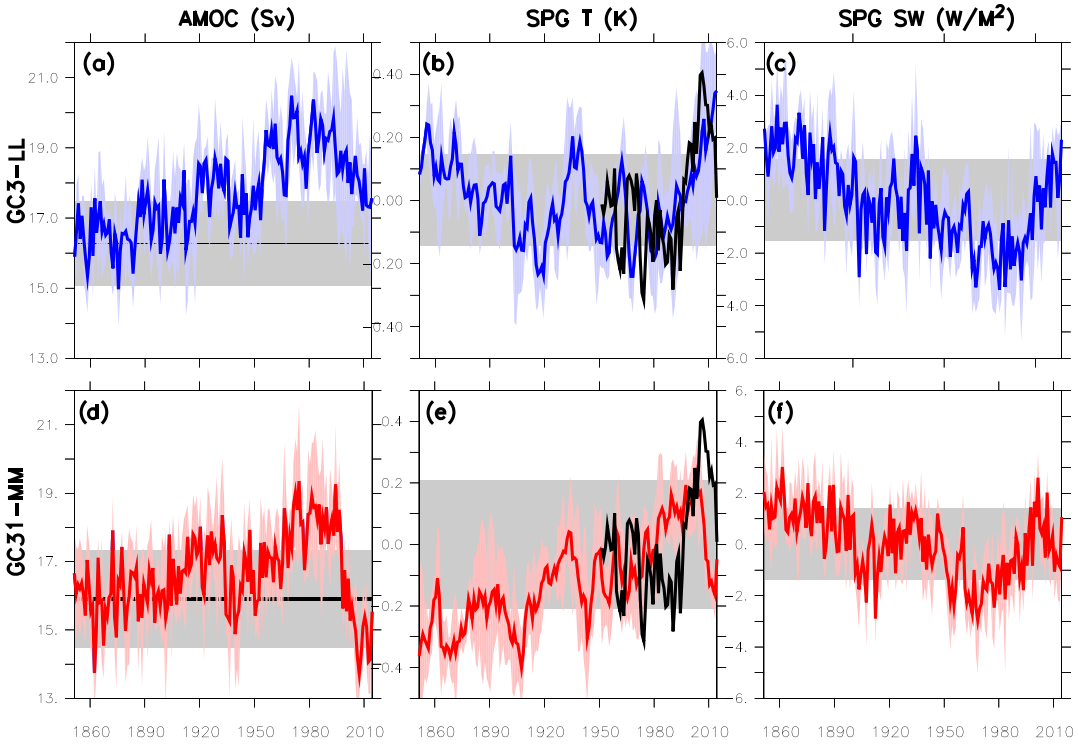

Figure 17. AMOC, subpolar heat content, and SW surface fluxes for GC31-LL (blue) and GC31-MM (red). (a) AMOC at 40°N (max of stream function—northward ocean velocity integrated from surface downward). Blue shading is 1 standard deviation of the ensemble spread. Black—mean of GC31-LL preindustrial (PI) control run (498 years) and shading is 1 standard deviation of the PI control. (b) GC31-LL ensemble mean ocean temperature anomaly (°C), mean over the Subpolar region (40:60°N, 0:60°W, 0–700 m). Shading as (a). Black: observations (EN4; Good et al., 2013). (c) GC31-LL Mean surface downward Shortwave flux (W/m^2) into the ocean surface over the Subpolar Gyre region. Shading as (a). (d–f) As in (a)–(c) but for the GC31-MM historical experiment and the GC31-MM *piControl* experiment (378 years).

2012). Therefore, as in previous studies, we find evidence that anthropogenic aerosol forcings have played a role in driving AMV (Booth et al., 2012; Watanabe & Tatebe, 2019) and AMOC (Delworth & Dixon, 2006; Menary et al., 2013; Undorf et al., 2018) in historical simulations. However, based on our analysis alone, we cannot rule out a role for natural forcing factors (Ottera et al., 2010). We note that, as with previous models (Zhang et al., 2013), HadGEM3-GC3.1 does not explain all the observed changes in AMV. However, we cannot rule out that these results are affected by the strong aerosol forcing in HadGEM3-GC3.1 (Bodas-Salcedo et al., 2019; Mulcahy et al., 2018), or a strong response to aerosol forcing, which may explain the strong correlation between Atlantic and Global forced responses in Figure 16. If the real world had a smaller role for aerosol forcing than that simulated in this model, then we would expect any forced response to have explained a smaller fraction of real-world variability. However, we also note that the comparisons with observations are subject to various uncertainties, including uncertainty in estimating the forced response due to the presence of internal variability in both models and observations and due to the uncertainty in producing the observations themselves.

5. Influence of Volcanic Eruptions on Surface Temperatures

Major tropical volcanic eruptions have a large impact on global climate for several years and should be included in historical experiments in order to capture an important component of subdecadal timescale climate variability (Smith et al., 2016). In this section, we attempt to quantify the impact of volcanic eruptions on global surface temperature. We use the approach of Driscoll et al. (2012) who looked at the impacts of volcanoes in CMIP5 models and include the five volcanoes that they used: Krakatau (erupted August 1883), Santa Maria (October 1902), Mount Agung (March 1963), El Chichón (April 1982), and Mount Pinatubo (June 1991). These are all large tropical volcanic eruptions. We avoid extratropical volcanoes as they appear to have a different impact on the circulation in the NH winter (Stenchikov et al., 2006). For each volcano,

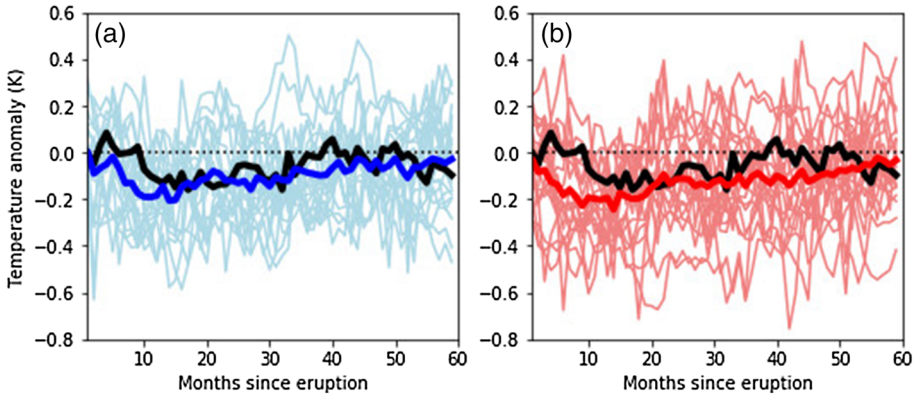

Figure 18. Global mean surface air temperature following major tropical volcanic eruptions for (a) GC31-LL, shown in blue, and (b) GC31-MM, in red. Each of the thin lines represents the temperature anomalies from one volcanic eruption and one member. The thick line is the average of these. The black line is the average of the five volcanoes in HadCRUT4. The model fields have been masked where there are no observations in HadCRUT4.

we calculate a climatology from the years preceding the volcanic eruption, the same approach as used in Driscoll et al. (2012), and subtract the climatology from the years following the volcanic eruption to obtain anomalies. These anomalies are a mixture of the impacts of the volcano and internal variability in the climate system, so we average over all volcanoes to reduce the influence of the latter.

Figure 18 shows how the global mean surface air temperature anomalies develop after a large volcanic eruption. In both GC31-LL and GC31-MM, the experiments cool, on average, approximately 0.2 Kafter 12 months. Subsequently, the temperature returns to close to normal about 4 years after the eruptions. There is a lot of variability across the range of ensemble members and volcanic eruptions, with some events not appearing to cool. The observed temperature anomalies from HadCRUT4 follow a very similar evolution to the models. The biggest difference with the model is in the first 12 months when models appear to cool more quickly, but given the large variability in model responses, this difference is likely due to chance. In conclusion, both GC31-LL and GC31-MM have realistic amounts of cooling in the global mean surface air temperature compared to HadCRUT4.

We now consider the pattern of surface temperature anomalies in the two NH winters following the eruptions. Past studies have shown that these feature counter-intuitive warming at high latitudes (Driscoll et al., 2012; Stenchikov et al., 2006) and in the tropical East Pacific (Khodri et al., 2017; Stevenson et al., 2017). Figure 19 shows surface temperature anomalies averaged over the first two winters (December-January-February) after the eruptions. As expected, there is widespread cooling in the tropics for both model resolutions and HadCRUT4. The exception is the tropical East Pacific, which has no significant anomalies in GC31-LL and some positive anomalies in HadCRUT4. If this tropical response is real and not due to chance in the observations, then these integrations do not capture it fully. In the extratropics, the main feature in HadCRUT4 is a warming greater than 1° over Scandinavia and northern Russia and a cooling of slightly less than a degree in the Middle East and South East Asia. A similar response is captured by both GC31-LL and GC31-MM, but the model anomalies are weaker and not as widespread. In conclusion, both GC31-LL and GC31-MM capture the patterns of the temperature response in the first two NH winters seen in HadCRUT4, but the magnitude of the anomalies is often weaker.

6. Influence of the SC on Atmospheric Temperatures and Precipitation

The 11-year SC influences the entire atmosphere throughout the historical period, either directly by absorption of spectral solar irradiances (SSI) or indirectly through complex coupling processes involving chemistry and dynamics (Gray et al., 2010; Seppala et al., 2014). While the total solar irradiance varies by merely ~1 W/m^2 (approximately 0.1% total solar irradiance) within the 11-year SC, the largest SSI variability is observed in the shorter wavelengths, causing significant stratospheric heating and ozone production (Haigh, 2013). This direct radiative-chemical interaction warms the upper stratosphere significantly and leads to subsequent dynamical changes which are proposed to transfer the solar signal to the surface, influencing the climate in the North Atlantic region (Andrews et al., 2015; Gray et al., 2013).

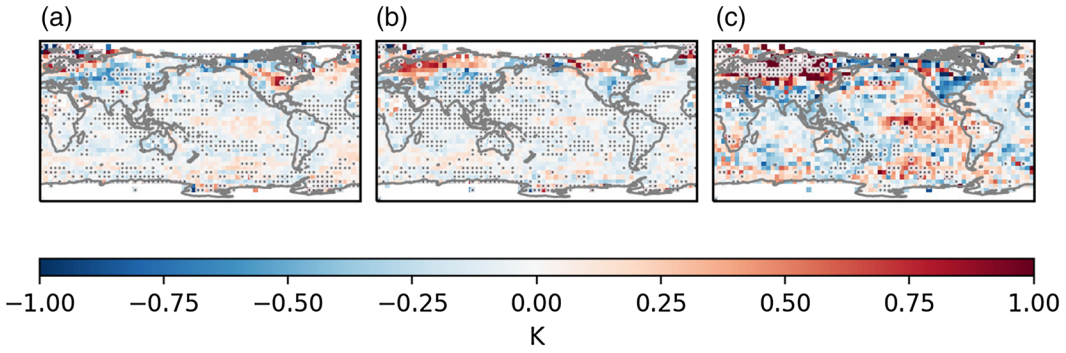

Figure 19. Surface temperature anomalies for the two Northern Hemisphere winters (December-January-February) following the volcanic eruptions for (a) GC31-LL, (b) GC31-MM, and (c) HadCRUT4. All major volcanic eruptions across all members have been averaged to bring out the robust response. Anomalies significant at the $p = 0.05$ level using a two-sided t test are stippled. The model fields have been masked where there are no observations in HadCRUT4.

We detect the SC contribution in the historical simulations with a multiple linear regression model that takes into account the SC, volcanic eruptions, El-Niño–Southern Oscillation, Quasi-Biennial Oscillation (Baldwin et al., 2001), and greenhouse gases as described in Misios et al. (2016). Given the periodic nature of the SC fluctuation, long-term background trends (insignificant in the stratosphere) do not bias the detection of solar signatures and hence we did not detrend the time series prior to analysis. The model calculates lagged SC regression coefficients by lagging the SC predictor by −5 to +5 years. The null hypothesis of zero regression coefficients is rejected with a P value smaller than 0.1 using a two-tailed t test.

Given that the same SSI and ozone concentrations are prescribed in both GC31-LL and GC31-MM, we expect very similar direct atmospheric heating in response to the SC in both model versions. Figure 20 shows this to be the case. The global atmospheric warming at solar maximum is about 0.7 K peaking in the upper stratosphere/lower mesosphere. There is no evidence of a time lag in the stratosphere and mesosphere, as anticipated by a direct heating. Analysis of zonal mean air temperature reveals a band of significant warming between 5 and 0.1 hPa, with a maximum in the upper tropical stratosphere of 0.8 K. The pattern of atmospheric warming shown in Figures 20b and 20d is consistent with the previous high-top version of the model (HadGEM2-CC) used for CMIP5 experiments (Mitchell et al., 2015), despite the different top-of-atmosphere (TOA) SSI and ozone concentrations that were specified.

The global mean atmospheric warming is considerably weaker in the troposphere. It does not exceed 0.2 K and is characterized by a time lag of 1 year (Figures 20a and 20c). A similar time lag is also found in the global mean surface temperature (not shown), presumably associated with the ocean absorption of the visible and infrared spectral range, leading to direct warming at the Earth's surface. This warming, however, is not distributed evenly over the Earth's surface because dynamical amplification through atmosphere-ocean interactions has the potential to amplify or dampen surface responses.

As shown in Figure 21, we find that the surface temperature response to the SC does not depend critically on the horizontal resolution given that both model configurations show surface patterns of stronger warming in the tropics. GC31-LL shows a warmer tropical Pacific and Atlantic Oceans with magnitude exceeding 0.2 K at lag +1 year. The surface response gets stronger in GC31-MM, where we find a broad warming over the Pacific taking a horseshoe shape in the North Pacific similar to the SC signature detected in the HadCRUT data set (Misios et al., 2016). Recent model simulations and analysis of historical precipitation records have identified a pattern of increased precipitation in the western Pacific at solar maximum (Misios et al., 2019). Figure 21 shows that a similar pattern of precipitation response is simulated at solar maximum. Both model configurations show increased precipitation of about 1 mm/day at lag +1 year in the equatorial Pacific. In GC31-LL, increased precipitation is found concentrated in the western Pacific, while GC31-MM shows increased rainfall over a broader region in the equatorial Pacific. This pattern of precipitation response is likely to be related to a slow down of the Walker Circulation associated with global changes in the hydrological cycle at solar maximum, but further analysis is needed to disentangle the actual mechanisms.

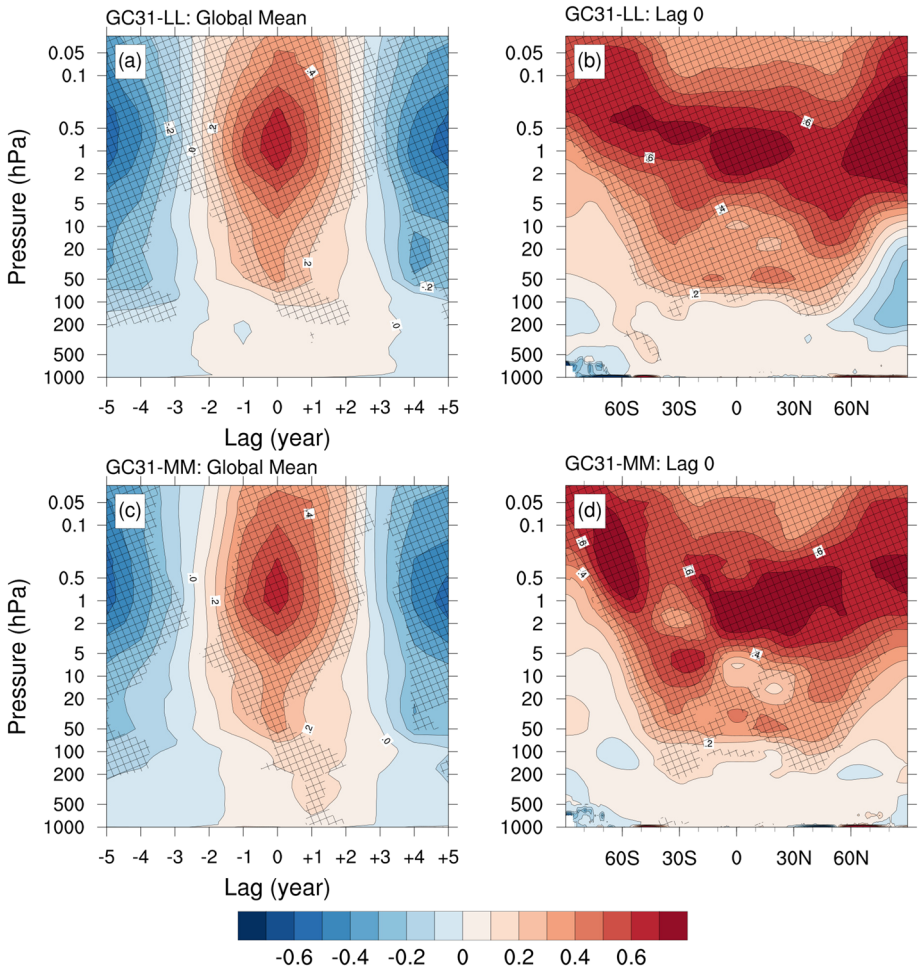

Figure 20. SC regression coefficients of annual global mean air temperature at different time lags (panels a, c) and zonal mean air temperature at lag 0 year (panels b, d). Units in K for a typical SC. Shading denotes statistical significance for p values ≤ 0.1.

7. Understanding the Model's Global Climate Response

In this section we use a simple energy balance framework to investigate plausible interpretations of the model's GMSAT time series (Figure 3). The model changes are within the envelope of the CMIP5 historical simulations, but we wish to develop a deeper understanding of what is driving the model GMSAT variations and their differences from observations. It is important to note that a successful simulation of the historical GMSAT time series does not guarantee that the model is correctly simulating the processes responsible for the historical changes (Golaz et al., 2013) nor that the model will correctly project the response of the climate system to future forcing (Gregory et al., 2019). Nonetheless, by identifying likely causes of model-observation differences, we develop insight into their likely impact on future projections and inform future process-based model development.

We focus on changes over certain periods to avoid the issue with choice of reference period mentioned in section 2.1. In particular, we wish to understand the weaker than observed warming from around 1930–1950, the stronger than observed cooling from around 1950–1980, and the stronger than observed warming from around 1980–2014.

Our analysis is based on the simple energy balance model:

$$N = F - \alpha T$$

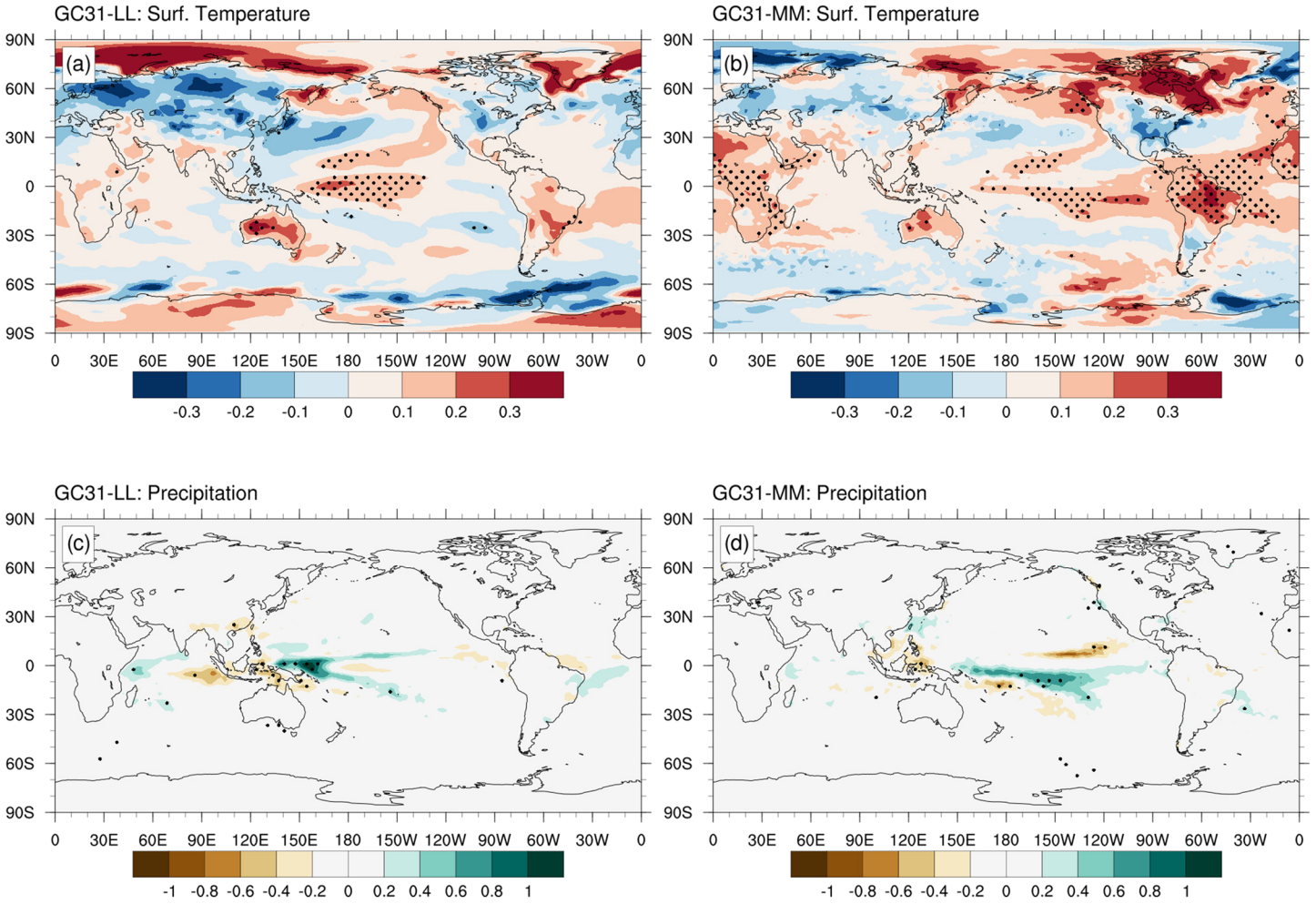

Figure 21. SC regression coefficients of annual mean surface air temperature (panels a, b) and total precipitation (panels c, d) at time lag +1 year. Units in K and mm/day for a typical SC. Shading denotes statistical significance for p values ≤ 0.1.

where $F(t)$ is the effective radiative forcing (ERF; Myhre et al., 2013), $N(t)$ is the TOA heat flux, $T(t)$ is the global mean surface temperature anomaly, and $\alpha(t)$ is a time- or state-dependent climate feedback parameter (inversely proportional to effective climate sensitivity). Taken over the whole period, there is a potential compensation between forcing and feedback. Specifically, a model with strong aerosol forcing (cooling) and high effective climate sensitivity could produce a similar temperature change over the whole historical period to a model with weak aerosol forcing and low effective climate sensitivity (Kiehl, 2007). By splitting the historical period into a few multidecadal periods, each characterized by different forcing trends, it may be possible to differentiate better between the drivers of the model-observation differences. A similar framework was used recently by (Golaz et al., 2019) to interpret the historical simulations in E3SM, another CMIP6 atmosphere-ocean general circulation model (AOGCM). Here we extend the approach of Golaz et al. (2019) by allowing α to vary with time. This is motivated by the analysis of Gregory et al. (2019) who show that α typically varies by a factor of 2 through the historical period in the CMIP5 historical runs. We ask in turn: Can the differences between observed and modeled GMSAT over the three periods be explained by differences in F, in α, or in internal climate variability, between the real world and the model?

7.1. ERF $F(t)$

ERF is a measure of the radiative forcing associated with particular climate forcing agents, including quasi-instantaneous feedbacks due to rapid adjustment of the atmosphere. Figure 22 shows the time series of F for the GC31-LL model, derived and described in Andrews et al. (2019) and the components due to well-mixed greenhouse gases and anthropogenic aerosols. These are compared with the partly observationally constrained central estimate of the IPCC Fifth Assessment Report (AR5 Myhre et al., 2013). We see

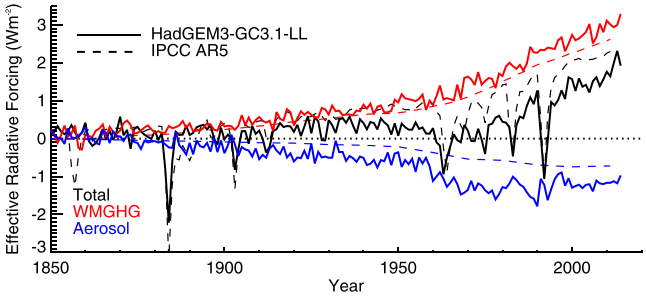

Figure 22. Effective radiative forcing $F(t)$ from well-mixed greenhouse gases (red), anthropogenic aerosols (blue), and total including all forcing agents (black). Solid lines show values evaluated for the GC31-LL model using the RFMIP *piClim-HISTall* run. Dashed lines show the central assessed estimate from IPCC AR5.

that the GC31-LL aerosol ERF is stronger than the AR5 central estimate throughout the 1900–2010 period. Uncertainty in the AR5 aerosol ERF estimate is large, with the assessed 5–95% range for 2010 relative to 1750 being [−0.1, −1.9] W/m^2. Recently, Bellouin et al. (2019) have reviewed post-AR5 evidence and propose a range for the period 1850–2010 of [−0.35, −2.0] W/m^2. The GC31-LL aerosol ERF of −1.1 W/m^2 (Andrews et al., 2019) is consistent with both of these ranges but several tenths of W/m^2 below (i.e., stronger than) the modal value of both. We are not aware of uncertainty estimates for the aerosol ERF at intermediate times within the twentieth century, but we note that the discrepancy with the AR5 central estimate is particularly large around 1950–1980. Note that the strong aerosol ERF in GC31-LL results in a total ERF that is decreasing between around 1930 and 1960. Because the TOA heat flux is always partially adjusted to the history of the ERF, the TOA flux can become zero during this period, even though the ERF is still positive relative to 1850. This accounts for the relatively constant OHC in the decades preceding 1960 (Figure 5).

Overall, we conclude that it seems likely (though by no means certain) that the modeled aerosol ERF is somewhat too strong through the twentieth century and especially in the 1950–1980 period. Such a bias would be consistent with the strong GMSAT cooling during that period.

7.2. Climate Feedback Parameter $\alpha(t)$

We calculated $\alpha(t)$ for the GC31-LL historical ensemble and for the corresponding *AMIP-piForcing* run (which has constant preindustrial forcing and prescribed SST from observations; see Webb et al., 2017), using the method of Gregory et al. (2019) (Figure 23). Both GC31-LL and *AMIP-piForcing* show a similar time variation of $\alpha(t)$ as seen for the CMIP5 models by Gregory et al. (2019) (their Figure 5c) but with overall values of α around 50% greater in GC31-LL than in CMIP5. The 30-year mean value of α at the end of the GC31-LL run is 0.8 W m^{-2} K^{-1}, while for the GC31-LL *AMIP-piForcing* run it is 1.8 W m^{-2} K^{-1}. Assuming an ERF due to doubled CO_2 concentration of 4.05 W/m^2 (Andrews et al., 2019), these correspond to effective climate sensitivities of 5.1°C and 2.2°C, respectively. The corresponding effective climate sensitivities for CMIP5 are 3.5°C and 1.7°C, respectively. Thus, we see that the apparently strong effective climate sensitivity in recent decades likely reflects both the known higher-than-average climate sensitivity of GC31 (Andrews et al., 2019; Bodas-Salcedo et al., 2019) and time variations in α driven by discrepancies between modeled and observed SST patterns in recent decades (Gregory et al., 2019), with the time variation being the dominant factor. We discuss the time variation further in the following paragraph.

Gregory et al. (2019) suggest that the lower recent values of α in the CMIP5 AOGCMs, compared with *AMIP-piForcing*, are due to the recent IPO-like observed SST pattern in the Pacific Ocean, which is associated with large α but is not seen in the AOGCMs. They propose that the discrepancy is either because the models' internal IPO variability is too low or because the SST response pattern to the volcanic forcing of the 1960s–1990s, which resembles the IPO pattern, is too weak. During the assessment of GC31's IPO variability (section 1.4), we found no evidence that the model variability was weaker than the (limited) observational record. However, the pattern of the short-term response to volcanic forcing is much weaker in the model than in the observations (see Figure 19, section 5). This suggests that for GC31, an underplayed volcanic response may be contributing to the underestimate of α in recent decades, as proposed for CMIP5 by Gregory et al. (2019). Further investigation using partially forced runs would be needed to confirm the cause but is beyond the scope of the present paper.

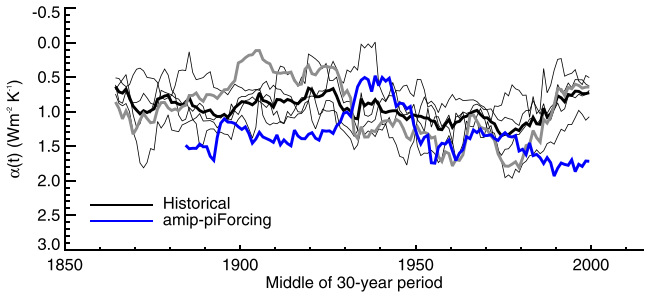

Figure 23. Climate feedback parameter $\alpha(t)$, evaluated by regressing T against F-N over a moving 30-year window, following Gregory et al. (2019). Blue: *AMIP-piForcing* experiment, providing an estimate of α given the observed SST field. Black: GC31-LL historical ensemble (individual members thin and their ensemble mean thick). Thick gray: as thick black but obtained by meaning the four members before performing the regression (Gregory et al., 2019). Note that the α axis is shown inverted for ease of comparison with Figure 5c of Gregory et al. (2019).

7.3. Internal Variability

The stronger than observed cooling between about 1950 and 1980 and stronger than observed warming between about 1980 and 2014 are seen in all ensemble members for GC31-LL and GC31-MM (eight members in all), and in all ensemble members of the UKESM1 Earth System Model, which uses GC31-LL as its physical climate core (Sellar et al., 2019) (16 further ensemble members). It is therefore unlikely that these features can be explained by internal climate variability, unless the models are underestimating the internal GMSAT variability of the real climate system.

In the *piControl* simulations of GC31-LL and GC31-MM we see interdecadal variations in GMSAT with detrended standard deviations of 0.055 and 0.062°C, respectively, similar in magnitude to the variations in the first part of the historical observed data set, before strong forced responses are expected (Figure 3). The corresponding figures for the NH only are 0.084°C and 0.088°C, which is comparable to the interdecadal variations seen in reconstructions of whole Northern Hemisphere surface temperature over the past 2,000 years (Masson-Delmotte et al., 2013). So there is no obvious evidence to suggest that the internal interdecadal variability of the model GMSAT is too low.

For the earlier period 1930–1950 a few members of the larger UKESM1 ensemble do appear to capture the warming from 1930–1950 ((Sellar et al., 2019), Figure 33), so internal variability is a plausible explanation of this difference.

7.4. Summary of Factors Affecting GMSAT Response

Based on our analysis above, we can draw some tentative conclusions about the causes of the discrepancies with the observed GMSAT time series:

1. The weaker-than-observed warming from around 1930–1950 in the GC31 historical ensembles is most likely due to some combination of strong ERF from anthropogenic aerosols and internal climate variability that is undersampled by our ensembles. We have been unable to separate these two causes in our analysis, but note that some members of the larger UKESM1 ensemble do capture the observed warming in this period. We cannot rule out a role for other forcing agents, for example, ozone which shows an accelerated increase during this period.
2. Internal variability is unlikely to explain the multidecadal differences between the models and observations since 1950. The most likely explanation of the strong cooling between 1950 and 1980 appears to be aerosol forcing that is stronger in the models than in the AR5 central estimate. However, we note the large uncertainty in the AR5 aerosol ERF estimate, as well as in more recent estimates.
3. In common with CMIP5 models (Gregory et al., 2019), the climate feedback parameter $\alpha(t)$ is too low (effective climate sensitivity too high) during the decades since 1980, when compared with estimates of α made using observed SSTs. The too low α is likely contributing to the strong warming since 1980. It seems most likely that a too muted SST pattern response to the volcanic eruptions of the 1960s–1990s is contributing to the erroneous SST patterns, but this would require further integrations to confirm. In any case the strong time dependence of α during this period suggests that warming over recent decades may

not be a good indicator of the response to substantially higher levels of greenhouse gases over the 21st century (Gregory et al., 2019).

4. The strong modeled rate of warming since 1990 is partly a result of the strong cooling spike following the 1991 Mount Pinatubo eruption (around 0.2°C stronger than observed). Around 0.1°C of this difference may be attributable to the moderate-to-strong El Niño observed in 1991–1992 (Folland et al., 2013).

Since the UKESM1 model (Sellar et al., 2019) uses GC31-LL as its physical climate core and its simulated GMSAT evolution is very similar to the runs studied here, we believe that our analysis applies also to UKESM1. Deeper study is needed to confirm our conclusions and is planned for a later paper.

8. Summary

We briefly summarize the key results arising from the analysis of the historical ensembles, GC31-LL, and GC31-MM. Various aspects of the simulated climate system have been assessed against available observations and shown to produce a reasonable representation of historical climate variability. The overall warming trends in the global mean surface air temperature between 1880–1891 and 2000–2014 are similar to observations; this is despite global mean land surface air temperature trends being smaller than observed. On average, the global temperature response to major tropical volcanic eruptions after 12 months exhibits realistic amounts of cooling, although internal variability complicates the comparison for individual eruptions. The pattern of warming in the high-latitude NH for the first two winters after these eruptions is captured, albeit weaker than observed. The observed low- to middle-latitude pattern of SST response to volcanic eruptions is weaker in the models, and this may have important implications for recent simulated rates of warming. The atmospheric response to SC variability shows a warming in the upper stratosphere-lower mesosphere at solar maximum consistent with the CMIP5 response. The two model resolutions exhibit similar temperature responses to the combination of historically evolving greenhouse gases, aerosols, volcanic eruptions, and solar variability.

The overall time series of global temperature in the historical runs resembles the observed time series but with exaggerated cooling from around 1930–1980 and stronger warming from 1980–2014. A simple analysis of these features suggests that a strong ERF from aerosols may be playing an important role in the cooling phase, while errors in the simulation of SST patterns since 1980 result in a too strong effective climate sensitivity during this period, possibly influenced by the above-mentioned weak SST response to the volcanic eruptions of the 1960s–1990s. Internal climate variability in the observations complicates the comparison before 1950 and after 1990. Work is in progress to develop a revised version of the model with improved representation of processes affecting aerosol forcing and will be reported elsewhere.

The historical ensembles generally do not tend to reproduce observed decadal mean precipitation in an ensemble mean sense, supporting the view that observed variability is internally generated and does not have much contribution from external forcings. Nonetheless, this conclusion is dependent on the fidelity of the forcings and responses in the models discussed here and others more generally. Precipitation variability is often mediated by dynamical changes (e.g., to monsoons), for which the sensitivity to climate forcings can be difficult to simulate. This is perhaps shown most clearly by the very large uncertainties in model projections of the 21st century regional precipitation, but the same differences in model processes could also affect the simulation of the 20th century. A common theme in the model ensembles is a tropical/subtropical drying from the 1950s to 1990s. This approximately coincides with the period when the models produce a cooling or hiatus of global mean temperatures. A resulting lack of evaporation and thereby lack of atmospheric moisture availability is consistent with the decline in rainfall. In some regions, observed rainfall variability is far in excess of the spread of the model's ensemble (e.g., West Africa), suggesting deficiencies in the model's ability to generate variability, or even perhaps that the forced response could actually be underestimated in the simulations. Also of note is the forced response to the 11-year SC that results in enhanced precipitation in the equatorial Pacific.

The OHC, an important indicator of the Earth's energy budget imbalance, shows a warming in the 0- to 700-m layer consistent with observations over the period 1971–2014. The lack of warming compared to observations in the 1970s–1980s is compensated by a warming rate larger than observed after 1993. The 0- to 2,000-m layer also exhibits no warming during the early period and warms at a similar rate to observations after 1993. This suggests that the post-1993 warming of the 0- to 700-m layer is too large and for the 700- to 2,000-m layer too little, compared to observations.

The Arctic summer sea ice extent for 1990–2009 is in agreement with observations for both resolutions of the model. Differences arise in the winter extent; the GC31-LL experiment is in good agreement with observations, whereas GC31-MM is underextensive in the Sea of Okhotsk and Labrador Sea. This may be linked to the differences in the location of deep convection and the path of the North Atlantic current in each model. In general, the GC31-MM Arctic winter sea ice extent is less extensive than in GC31-LL over the entire historical period but is within near-present-day observational uncertainties. The negative trends in the near-present-day sea ice extent are reasonably consistent with observations at both resolutions.

For the Antarctic sea ice, both resolutions are underextensive in winter and summer for 1990–2009, more so for GC31-MM than for GC31-LL. This is linked to the Southern Ocean warm bias present in both models, which is greater for GC31-MM than for GC31-LL. The slightly positive trends in observations of the near-present-day sea ice extent are not reflected by the experiments.

The magnitude of the SMB of the Greenland and Antarctic ice sheets in the GC31-MM historical experiment is consistent with observations, suggesting that the constant SMB of the Antarctic ice sheet and the near-present-day decline in the Greenland ice sheet is captured by the higher resolution model, despite the model lacking ice sheet dynamical processes. The GC31-LL experiment exhibits a smaller SMB than observations, which may be linked to weaker coastal orographic snowfall at low resolution, but still manages to capture the magnitude of the near-present-day decline in the Greenland ice sheet. The area of permafrost, which is related to air temperature, in the low and medium resolution models is slightly larger than for observations. This is likely due to a colder Arctic in the model. The near-present-day permafrost loss rate is approximately double that of observations. Conversely, the loss of snow cover is lower than but within an order of magnitude of observations.

The principal mode of variability of North Atlantic SSTs, the AMV, appears to be associated with trends in the global mean surface air temperature in the historical simulations, unlike for the observed AMV. When trends are removed, the correlation between the model and observed AMV is greatly reduced. However, regression of North Atlantic SSTs with the AMV, having removed the influence of global mean SSTs, indicates that a significant proportion of the variance in the tropical Atlantic can be explained by forcing, that is, they are consistent across the historical ensembles. This is consistent with observations, supporting the idea that the observed tropical AMV has a forced component. It also suggests that internal variability is more important for the extratropical AMV. The evolution of the AMOC on decadal timescales is strongly correlated between the low and medium resolution experiments, exhibiting an increasing circulation up to 1,990 followed by a decline. This correlation suggests that AMOC variability is somehow being forced, possibly linked to upper ocean temperatures and shortwave surface fluxes. The variations in net shortwave radiation are linked to sulfate aerosol precursor emissions (Booth et al., 2012), providing a potential link between anthropogenic aerosol forcings and variability in the AMV and possibly AMOC.

Is model resolution an important factor when interpreting the results of the historical experiments? The answer differs depending upon which phenomenon is under scrutiny. The response of the GC31-LL and GC31-MM experiments is similar in some cases and divergent in others. For example, phenomena that have a direct influence on atmospheric and surface air temperatures, such as major volcanic eruptions, solar variability, greenhouse gas concentrations, and aerosols, appear to invoke similar responses at each resolution. Equally, the similar precipitation response of both ensembles suggests that the models are responding to external forcings in a consistent way. It should be noted that the small number of ensemble members, four at each resolution, may influence this interpretation. It is possible that larger ensembles may reveal significant differences between the resolutions.

Model resolution appears to be an important consideration for sea ice extent, reflecting known differences in the mean sea ice state in the model *piControl* runs. In the Arctic winter, the extent of the sea ice is greater in the lower resolution experiment and is suspected to be influenced by the different centers of deep circulation in northern oceans. In the Antarctic, for both winter and summer, the sea ice is underextensive for GC31-MM compared to GC31-LL and is thought to be related to a stronger Southern Ocean warm bias in GC31-MM.

In conclusion, the results of the historical simulations performed using the HadGEM3-GC3.1 model at N96/ORCA1 and N216/ORCA025 resolutions show many features in common with the observed record over the period 1850–2014. The simulations contribute, as part of the wider CMIP6 multimodel effort, to

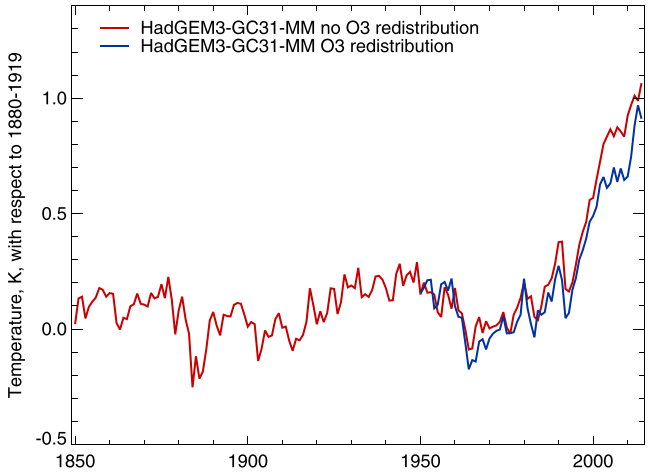

Figure A1. Global mean surface air temperature (GMSAT) evolution for GC31-MM historical experiments with (blue) and without (red) ozone redistribution. Shaded region indicate ±2 standard deviations. The *piControl* trend over 500 years was removed, and the spatial anomalies found with respect to 1961–1990, prior to anomalizing to the 1880–1919 period. No observational masking has been applied.

understanding the causes of observed climate change since 1850. Future in-depth analysis of the mechanisms of change in models and observations will lead to an improved understanding of how the historical record can be used to better constrain model projections of future climate.

Appendix A: Ozone Redistribution in GC31-MM Historical Experiments From 1951 to 2014

The physical HadGEM3-GC3.1 CMIP6 model experiments use prescribed monthly ozone concentrations derived from the CMIP6 ozone concentration data sets as detailed in Eyring et al. (2016). While the climate sensitivity experiments were running, it was discovered that the height variation of the ozone tropopause, implicit in the prescribed ozone data set, did not match the changing tropopause height in the experiments. This was also found to occur, to a much lesser extent, in the near-present-day portion of the historical experiments. This mismatch results in increased concentrations of stratospheric water vapor which has a radiative influence on the surface climate sensitivity. The solution was the implementation of an ozone redistribution scheme within the model which acts to align the prescribed ozone tropopause with that of the running experiment. Hardiman et al. (2019) describe the implementation of this scheme. The inclusion of the scheme in the GC31-LL *piControl* simulation was shown by Hardiman et al. (2019) to have negligible impact, whereas the impact on the *abrupt-4xCO2* experiment was significant due to increasing tropopause height as the climate warmed. The historical simulations are more closely aligned with the *piControl* than the more extreme *abrupt-4xCO2* experiments, with only a small increase in tropopause height occurring toward the end of the historical period.

The GC31-LL historical simulations described in this paper use the ozone redistribution scheme throughout, from 1850 to 2014. The computationally intensive GC31-MM historical simulations could not be restarted from 1850 with this scheme due to CMIP6 delivery timescales. Hence, the ozone redistribution scheme was introduced partway through the simulation, from 1 January 1950, becoming effective 1 January 1951, over a decade prior to the impact of the Mount Agung volcanic eruption and a few decades prior to the near-present-day increase in global mean temperature.

Here we provide a brief impact assessment of introducing the ozone redistribution scheme partway through the GC31-MM historical simulation. The evolution of the ensemble mean global mean near surface temperature is shown in Figure A1. The original historical ensemble without ozone redistribution (1850–2014) is shown in red and the historical ensemble with ozone redistribution (1950–2014) in blue. The ensemble members of the latter are initialized using 1 January 1950 initial conditions generated by the nonozone redistributed experiments. For clarity, the DECK historical ensemble delivered for CMIP6 is a combination of the red curve prior to 1950 and the blue curve from 1950 onward.

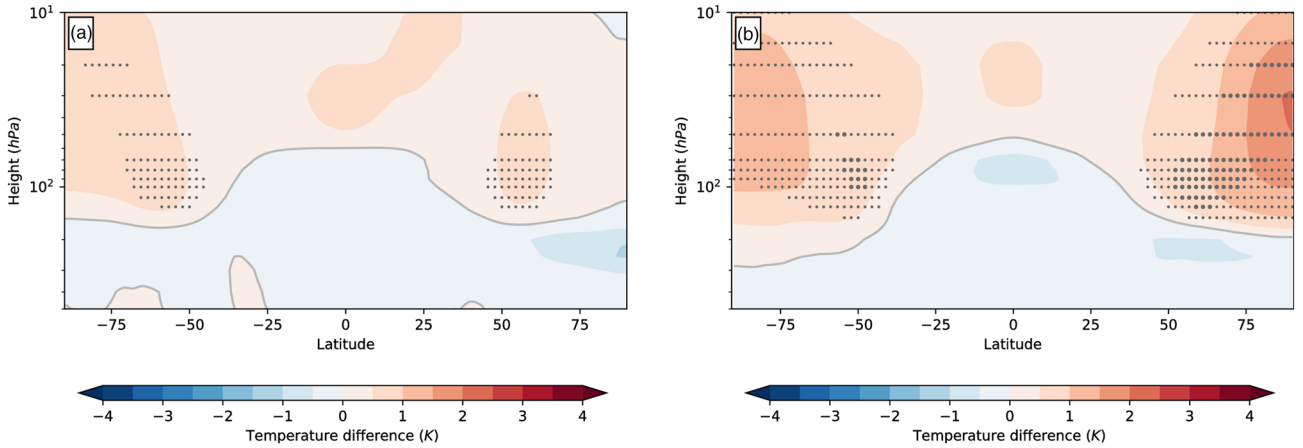

Figure A2. Zonal mean temperature difference between GC31-MM historical ensembles with and without ozone redistribution for the periods (a) 1951–1955 and (b) 2010–2014. Significance is calculated against the standard deviation of the 1850–2014 annual zonal mean temperature. Small stippling denotes one sigma, large stippling denotes two sigma.

The numerical integration of the two ensembles diverges due to the influence of the ozone redistribution scheme, all other factors being equal. Comparing the ensemble means of both experiments, the DECK experiment warms slightly prior to 1959 and thereafter remains cooler, although these differences are not significant, which is a result consistent with the *piControl* results described in Hardiman et al. (2019). A comparison of the mean difference over the period 1951–1955 (−0.01 K) with that of 2010–2014 (+0.16 K) suggests that the influence of ozone redistribution becomes increasingly important toward the end of the historical period, as anticipated.

We also consider the influence of the ozone redistribution scheme on the temperatures in the troposphere and stratosphere. The zonal mean zonal temperature differences between the original and the DECK historical experiments over the periods 1951–1955, and 2010–2014, shown in Figure A2, are attributable to a combination of the ozone redistribution scheme and internal variability associated with the numerical divergence of the two ensembles. A small impact on ozone concentration and zonal mean temperature is inevitable in the first few years after the scheme is introduced (Figure A2a), but is not significant for zonal mean zonal temperature in the context of interannual variability. Comparison with the equivalent Figure 7a of Hardiman et al. (2019), demonstrates that the response in the early 1950s is commensurate with that of the *piControl* experiment. There is more significant impact in the near-present day (Figure A2b), with a larger stratospheric warming response, although this is still weaker than the *abrupt-4xCO2* experiment (see Figure 7b of Hardiman et al., 2019). This demonstrates that the "correction" provided by the ozone redistribution scheme is working as intended.

It should be noted that the 1950–2014 GC31-MM historical experiment *without* the ozone redistribution scheme, which was used by the CMIP6 Decadal Climate Prediction Project (DCPP), does not form part of the UK-CMIP6 DECK contribution. It is likely that the 1960 onward GC31-MM historical ensemble without ozone redistribution will be made available as part of DCPP.

Acknowledgments
M. B. A, E. B., S. C. H, L. H., M. A. R, and G. J. S. were supported by the Met Office Hadley Centre Climate Programme funded by BEIS and Defra. E. B., T. K., and M. S. M. were supported by EU Horizon 2020 CRESCENDO Project, Grant 641816. L. H. was supported by the Climate Resilience Project funded by UKRI. M. S. M. was supported by the Newton Fund through the Met Office Climate Science for Service Partnership Brazil (CSSP Brazil). T. K. was funded by the National Environmental Research Council (NERC) national capability grant for the UK Earth System Modeling project, Grant NE/N017951/1. We thank Jonathan Gregory for valuable discussion of the time variation of climate feedback.

Data Availability. The simulation data used in this study are archived on the Earth System Grid Federation (ESGF) node (https://esgf-index1.ceda.ac.uk/). The model Source IDs are HadGEM3-GC31-LL (Ridley et al., 2019a) and HadGEM3-GC31-MM (Ridley et al., 2019b). The four members of each historical ensemble are identified by the following Variant Labels: r1i1p1f3 to r4i1p1f3.

References

Agosta, C., Amory, C., Kittel, C., Orsi, A., Favier, V., Gallée, H., et al. (2019). Estimation of the Antarctic surface mass balance using the regional climate model MAR (1979–2015) and identification of dominant processes. *The Cryosphere*, *13*(1), 281–296. https://doi.org/10.5194/tc-13-281-2019

Andrews, T., Andrews, M. B., Bodas-Salcedo, A., Jones, G. S., Kuhlbrodt, T., Manners, J., et al. (2019). Forcings, feedbacks and climate sensitivity in HadGEM3GC3.1 and UKESM1. *Journal of Advances in Modeling Earth Systems*, *11*, 4377–4394. https://doi.org/10.1029/2019MS001866

Andrews, M. B., Knight, J. R., & Gray, L. J. (2015). A simulated lagged response of the North Atlantic Oscillation to the solar cycle over the period 1960–2009. *Environmental Research Letters*, *10*(5), 054,022.

Baldwin, M., Gray, L., Dunkerton, T., Hamilton, K., Haynes, P., Randel, W., et al. (2001). The quasi-biennial oscillation. *Reviews of Geophysics*, *39*(2), 179–229.

Becker, A., Finger, P., Meyer-Christoffer, A., Rudolf, B., Schamm, K., Schneider, U., & Ziese, M. (2013). A description of the global land-surface precipitation data products of the Global Precipitation Climatology Centre with sample applications including centennial (trend) analysis from 1901-present. *Earth System Science Data*, *5*(1), 71–99. https://doi.org/10.5194/essd-5-71-2013

Bellouin, N., Quaas, J., Gryspeerdt, E., Kinne, S., Stier, P., Watson-Parris, D., et al. (2019). Bounding global aerosol radiative forcing of climate change. *Reviews of Geophysics*, *58*, e2019RG000660. https://doi.org/10.1029/2019RG000660

Biasutti, M. (2013). Forced Sahel rainfall trends in the CMIP5 archive. *Journal of Geophysical Research: Atmospheres*, *118*, 1613–1623. https://doi.org/10.1002/jgrd.50206

Bodas-Salcedo, A., Mulcahy, J. P., Andrews, T., Williams, K. D., Ringer, M. A., Field, P. R., & Elsaesser, G. S. (2019). Strong dependence of atmospheric feedbacks on mixed-phase microphysics and aerosol-cloud interactions in HadGEM3. *Journal of Advances in Modeling Earth Systems*, *11*, 1735–1758. https://doi.org/10.1029/2019MS001688

Booth, B. B., Dunstone, N. J., Halloran, P. R., Andrews, T., & Bellouin, N. (2012). Aerosols implicated as a prime driver of twentieth-century North Atlantic climate variability. *Nature*, *484*(7393), 228.

Chadburn, S. E., Burke, E., Cox, P., Friedlingstein, P., Hugelius, G., & Westermann, S. (2017). An observation-based constraint on permafrost loss as a function of global warming. *Nature Climate Change*, *7*(5), 340.

Chang, C.-Y., Chiang, J. C. H., Wehner, M. F., Friedman, A. R., & Ruedy, R. (2010). Sulfate aerosol control of tropical Atlantic climate over the twentieth century. *Journal of Climate*, *24*(10), 2540–2555. https://doi.org/10.1175/2010JCLI4065.1

Cheng, L., Trenberth, K. E., Fasullo, J., Boyer, T., Abraham, J., & Zhu, J. (2017). Improved estimates of ocean heat content from 1960 to 2015. *Science Advances*, *3*(3), e1601545. https://doi.org/10.1126/sciadv.1601545

Delworth, T. L., & Dixon, K. W. (2006). Have anthropogenic aerosols delayed a greenhouse gas-induced weakening of the North Atlantic thermohaline circulation? *Geophysical Research Letters*, *33*, L02606. https://doi.org/10.1029/2005GL024980

Domingues, C. M., Church, J. A., White, N. J., Gleckler, P. J., Wijffels, S. E., Barker, P. M., & Dunn, J. R. (2008). Improved estimates of upper-ocean warming and multi-decadal sea-level rise. *Nature*, *453*, 1090–1093. https://doi.org/10.1038/nature07080

Driscoll, S., Bozzo, A., Gray, L. J., Robock, A., & Stenchikov, G. (2012). Coupled Model Intercomparison Project 5 (CMIP5) simulations of climate following volcanic eruptions. *Journal of Geophysical Research*, *117*, D17105. https://doi.org/10.1029/2012JD017607

Estilow, T. W., Young, A. H., & Robinson, D. A. (2015). A long-term Northern Hemisphere snow cover extent data record for climate studies and monitoring. *Earth System Science Data*, *7*(1), 137–142.

Ettema, J., van den Broeke, M. R., van Meijgaard, E., van de Berg, W. J., Bamber, J. L., Box, J. E., & Bales, R. C. (2009). Higher surface mass balance of the Greenland ice sheet revealed by high-resolution climate modeling. *Geophysical Research Letters*, *36*, L12501. https://doi.org/10.1029/2009GL038110

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., & Taylor, K. E. (2016). Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geoscientific Model Development*, *9*(5), 1937–1958. https://doi.org/10.5194/gmd-9-1937-2016

Fetterer, F., Knowles, K., Meier, W., Savoie, M., & Windnagel, A. (2017). updated daily: Sea ice index, version 3. National snow and ice data center . [Date Accessed: 15 Nov 2019].

Flato, G., Marotzke, J., Abiodun, B., Braconnot, P., Chou, S. C., Collins, W., et al. (2013). Evaluation of climate models. In Stocker, T. F., Qin, D., Plattner, G.-K., Tignor, M., Allen, S. K., Doschung, J., Nauels, A., Xia, Y., Bex, V., & Midgley, P. M. (Eds.), *Climate change 2013: The physical science basis. Contribution of working group I to the fifth assessment report of the intergovernmental panel on climate change* pp. 741–882). Cambridge, UK: Cambridge University Press.

Folland, C. K., Colman, A. W., Smith, D. M., Boucher, O., Parker, D. E., & Vernier, J. P. (2013). High predictive skill of global surface temperature a year ahead. *Geophysical Research Letters*, *40*, 761–767. https://doi.org/10.1002/grl.50169

Folland, C. K., Palmer, T. N., & Parker, D. E. (1986). Sahel rainfall and worldwide sea temperatures, 1901–85. *Nature*, *320*(6063), 602.

Golaz, J.-C., Caldwell, P. M., Van Roekel, L. P., Petersen, M. R., Tang, Q., Wolfe, J. D., et al. (2019). The DOE E3SM coupled model version 1: Overview and evaluation at standard resolution. *Journal of Advances in Modeling Earth Systems*, *11*, 2089–2129. https://doi.org/10.1029/2018MS001603

Golaz, J.-C., Horowitz, L. W., & Levy II, H. (2013). Cloud tuning in a coupled climate model: Impact on 20th century warming. *Geophysical Research Letters*, *40*, 2246–2251. https://doi.org/10.1002/grl.50232

Good, S. A., Martin, M. J., & Rayner, N. A. (2013). EN4: Quality controlled ocean temperature and salinity profiles and monthly objective analyses with uncertainty estimates. *Journal of Geophysical Research: Oceans*, *118*, 6704–6716. https://doi.org/10.1002/2013JC009067

Gray, L., Beer, J., Geller, M., Haigh, J. D., Lockwood, M., Matthes, K., et al. (2010). Solar influences on climate. *Reviews of Geophysics*, *48*, RG4001. https://doi.org/10.1029/2009RG000282

Gray, L., Scaife, A. A., Mitchell, D. M., Osprey, S., Ineson, S., Hardiman, S., et al. (2013). A lagged response to the 11 year solar cycle in observed winter Atlantic/European weather patterns. *Journal of Geophysical Research: Atmospheres*, *118*, 13,405–13,420. https://doi.org/10.1002/2013JD020062

Gregory, J. M., Andrews, T., Ceppi, P., Mauritsen, T., & Webb, M. J. (2019). How accurately can the climate sensitivity to CO_2 be estimated from historical climate change? *Climate Dynamics*, *54*, 129–157.

Haigh, J. D. (2013). Solar variability and the stratosphere, *The stratosphere: Dynamics, transport, and chemistry* (pp. 173–187): American Geophysical Union. https://doi.org/10.1002/9781118666630.ch10

Hardiman, S. C., Andrews, M. B., Andrews, T., Bushell, A. C., Dunstone, N. J., Dyson, H., et al. (2019). The impact of prescribed ozone in climate projections run with HadGEM3-GC3.1. *Journal of Advances in Modeling Earth Systems*, *11*, 3443–3453. https://doi.org/10.1029/2019MS001714

Hartmann, D. L., Tank, A. M. K., Rusticucci, M., Alexander, L. V., Brönnimann, S., Charabi, Y. A. R., et al. (2013). Observations: Atmosphere and surface, *Climate change 2013 the physical science basis: Working group I contribution to the fifth assessment report of the intergovernmental panel on climate change* pp. 159–254): Cambridge University Press.

Hawkins, E., & Sutton, R. (2009). Connecting climate model projections of global temperature change with the real world. *Bulletin of the American Meteorological Society*, *97*(6), 963–980. https://doi.org/10.1175/BAMS-D-14-00154.1

Hunke, E. C., Lipscomb, W. H., Turner, A. K., Jeffery, N., & Elliott, S. (2015). CICE: The Los Alamos sea ice model documentation and software user's manual version 5.1 LA-CC-06-012. Los Alamos National Laboratory.

Hyder, P., Edwards, J. M., Allan, R. P., Hewitt, H. T., Bracegirdle, T. J., Gregory, J. M., et al. (2018). Critical Southern Ocean climate model biases traced to atmospheric model cloud errors. *Nature communications*, *9*, 3625.

Ishii, M., Fukuda, Y., Hirahara, S., Yasui, S., Suzuki, T., & Sato, K. (2017). Accuracy of global upper ocean heat content estimation expected from present observational data sets. *SOLA*, *13*, 163–167.

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

Jones, G. S., Stott, P. A., & Christidis, N. (2013). Attribution of observed historical near surface temperature variations to anthropogenic and natural causes using CMIP5 simulations. *Journal of Geophysical Research: Atmospheres*, *118*, 4001–4024. https://doi.org/10.1002/jgrd.50239

Kerr, R. A. (2000). A North Atlantic climate pacemaker for the centuries. *Science*, *288*(5473), 1984–1985. https://doi.org/10.1126/science.288.5473.1984

Khodri, M., Izumo, T., Vialard, J., Janicot, S., Cassou, C., Lengaigne, M., et al. (2017). Tropical explosive volcanic eruptions can trigger El Niño by cooling tropical Africa. *Nature Communications*, *8*, 778. https://doi.org/10.1038/s41467-017-00755-6

Kiehl, J. T. (2007). Twentieth century climate model response and climate sensitivity. *Geophysical Research Letters*, *34*, L22710. https://doi.org/10.1029/2007GL031383

Knight, J. R., Folland, C. K., & Scaife, A. A. (2006). Climate impacts of the Atlantic multidecadal oscillation. *Geophysical Research Letters*, *33*, 3867–3876. https://doi.org/10.1002/2017GL072681

Kuhlbrodt, T., Jones, C. G., Sellar, A., Storkey, D., Blockley, E., Stringer, M., et al. (2018). The low-resolution version of HadGEM3 GC3.1: Development and evaluation for global climate. *Journal of Advances in Modeling Earth Systems*, *10*, 2865–2888. https://doi.org/10.1029/2018MS001370

Lenaerts, J. T. M., Medley, B., van den Broeke, M. R., & Wouters, B. (2019). Observing and modeling ice sheet surface mass balance. *Reviews of Geophysics*, *57*, 376–420. https://doi.org/10.1029/2018RG000622

Lenaerts, J. T. M., van den Broeke, M. R., van de Berg, W. J., van Meijgaard, E., & Kuipers Munneke, P. (2012). A new, high-resolution surface mass balance map of Antarctica (1979–2010) based on regional atmospheric climate modeling. *Geophysical Research Letters*, *39*, L04501. https://doi.org/10.1029/2011GL050713

Levitus, S., Antonov, J. I., Boyer, T. P., Baranova, O. K., Garcia, H. E., Locarnini, R. A., et al. (2012). World ocean heat content and thermosteric sea level change (0-2000 m), 1955-2010. *Geophysical Research Letters*, *39*, L10603. https://doi.org/10.1029/2012GL051106

Madec, G., Bourdall-Badie, R., Bouttier, P.-A., Bricaud, C., Bruciaferri, D., Calvert, D., et al. (2017). NEMO ocean engine. Zenodo.

Masson-Delmotte, V., Schulz, M., Abe-Ouci, A., Beer, J., Ganopolski, A., GonzalezRouco, J. F., et al. (2013). Information from palaeoclimate archives, *Climate change 2013: The physical science basis. Contribution of working group I to the fifth assessment report of the intergovernmental panel on climate change* (pp. 383–464). Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/CBO9781107415324.018

Matthes, K., Funke, B., Andersson, M. E., Barnard, L., Beer, J., Charbonneau, P., et al. (2017). Solar forcing for CMIP6 (v3.2). *Geoscientific Model Development*, *10*(6), 2247–2302. https://doi.org/10.5194/gmd-10-2247-2017

Menary, M. B., Kuhlbrodt, T., Ridley, J., Andrews, M. B., Dimdore-Miles, O. B., Deshayes, J., et al. (2018). Preindustrial control simulations with HadGEM3-GC3.1 for CMIP6. *Journal of Advances in Modeling Earth Systems*, *10*, 3049–3075. https://doi.org/10.1029/2018MS001495

Menary, M. B., Roberts, C. D., Palmer, M. D., Halloran, P. R., Jackson, L., Wood, R. A., et al. (2013). Mechanisms of aerosol-forced AMOC variability in a state of the art climate model. *Journal of Geophysical Research: Oceans*, *118*, 2087–2096. https://doi.org/10.1002/jgrc.20178

Misios, S., Gray, L. J., Knudsen, M. F., Karoff, C., Schmidt, H., & Haigh, J. D. (2019). Slowdown of the Walker circulation at solar cycle maximum. *Proceedings of the National Academy of Sciences*, *116*(15), 7186–7191. https://doi.org/10.1073/pnas.1815060116

Misios, S., Mitchell, D. M., Gray, L. J., Tourpali, K., Matthes, K., Hood, L., et al. (2016). Solar signals in CMIP-5 simulations: Effects of atmosphere-ocean coupling. *Quarterly Journal of the Royal Meteorological Society*, *142*, 928–941.

Mitchell, D. M., Misios, S., Gray, L. J., Tourpali, K., Matthes, K., Hood, L., et al. (2015). Solar signals in CMIP-5 simulations: The stratospheric pathway. *Quarterly Journal of the Royal Meteorological Society*, *141*, 2390–2403.

Monerie, P.-A., Robson, J., Dong, B., Hodson, D. L. R., & Klingaman, N. P. (2019). Effect of the Atlantic multidecadal variability on the global monsoon. *Geophysical Research Letters*, *46*, 1765–1775. https://doi.org/10.1029/2018GL080903

Morice, C. P., Kennedy, J. J., Rayner, N. A., & Jones, P. D. (2012). Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: The HadCRUT4 dataset. *Journal of Geophysical Research*, *117*, D08101. https://doi.org/10.1029/2011JD017187

Mouginot, J., Rignot, E., Bjørk, A. A., van den Broeke, M., Millan, R., Morlighem, M., et al. (2019). Forty-six years of Greenland Ice Sheet mass balance from 1972 to 2018. *Proceedings of the National Academy of Sciences*, *116*(19), 9239–9244. https://doi.org/10.1073/pnas.1904242116

Mulcahy, J. P., Jones, C., Sellar, A., Johnson, B., Boutle, I. A., Jones, A., et al. (2018). Improved aerosol processes and effective radiative forcing in HadGEM3 and UKESM1. *Journal of Advances in Modeling Earth Systems*, *10*, 2786–2805. https://doi.org/10.1029/2018MS001464

Myhre, G., Shindell, D., Bron, F.-M., Collins, W., Fuglestvedt, J., Huang, J., et al. (2013). Anthropogenic and natural radiative forcing. In Stocker, T. F., Qin, D., Plattner, G.-K., Tignor, M., Allen, S. K., Doschung, J., Nauels, A., Xia, Y., Bex, V., & Midgley, P. M. (Eds.), *Climate change 2013: The physical science basis. Contribution of working group I to the fifth assessment report of the intergovernmental panel on climate change* (pp. 659–740). Cambridge, UK: Cambridge University Press.

Ottera, O. H., Bentsen, M., Drange, H., & Suo, L. (2010). External forcing as a metronome for Atlantic multidecadal variability. *Nature Geoscience*, *3*(10), 688–694. https://doi.org/10.1038/ngeo955

Power, S., Casey, T., Folland, C., Colman, A., & Mehta, V. (1999). Inter-decadal modulation of the impact of ENSO on Australia. *Climate Dynamics*, *15*(5), 319–324.

Qasmi, S., Cassou, C., & Bo, J. (2017). Teleconnection between Atlantic multidecadal variability and European temperature: Diversity and evaluation of the coupled model intercomparison project phase 5 models. *Geophysical Research Letters*, *44*, 11,140–11,149. https://doi.org/10.1002/2017GL074886

Rayner, N. A., Parker, D. E., Horton, E. B., Folland, C. K., Alexander, L. V., Rowell, D. P., et al. (2003). Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. *Journal of Geophysical Research*, *108*(D14), 4407. https://doi.org/10.1029/2002JD002670

Ridley, J. K., Blockley, E. W., Keen, A. B., Rae, J. G. L., West, A. E., & Schroeder, D. (2018). The sea ice model component of HadGEM3-GC3.1. *Geoscientific Model Development*, *11*(2), 713–723. https://doi.org/10.5194/gmd-11-713-2018

Ridley, J. K., Menary, M., Kuhlbrodt, T., Andrews, M., & Andrews, T. (2019). MOHC HadGEM3-GC31-LL model output prepared for CMIP6 CMIP historical. https://doi.org/10.22033/ESGF/CMIP6.6109

Ridley, J. K., Menary, M., Kuhlbrodt, T., Andrews, M., & Andrews, T. (2019). MOHC HadGEM3-GC31-MM model output prepared for CMIP6 CMIP historical. https://doi.org/10.22033/ESGF/CMIP6.6112

Roberts, M. J., Baker, A., Blockley, E. W., Calvert, D., Coward, A., Hewitt, H. T., et al. (2019). Description of the resolution hierarchy of the global coupled HadGEM3-GC3.1 model as used in CMIP6 HighResMIP experiments. *Geoscientific Model Development*, *12*(12), 4999–5028. https://doi.org/10.5194/gmd-12-4999-2019

Roberts, C. D., Jackson, L., & McNeall, D. (2014). Is the 2004–2012 reduction of the Atlantic meridional overturning circulation significant? *Geophysical Research Letters*, *41*, 3204–3210. https://doi.org/10.1002/2014GL059473

Robinson, D. A., & Mote, T. L. (2014). MEaSUREs Northern Hemisphere terrestrial snow cover extent weekly 100km EASE-Grid 2.0, version 1. Boulder, Colorado USA. NASA National Snow and Ice Data Center Distributed Active Archive Center. https://doi.org/10.5067/MEASURES/CRYOSPHERE/nsidc-0531.001

Robson, J., Ortega, P., & Sutton, R. (2016). A reversal of climatic trends in the North Atlantic since 2005. *Nature Geoscience*, *9*(7), 513–517. https://doi.org/10.1038/ngeo2727

Scaife, A., Kucharski, F., Folland, C., Kinter, J., Brönnimann, S., Fereday, D., et al. (2009). The CLIVAR C20C project: Selected twentieth century climate events. *Climate Dynamics*, *33*(5), 603–614.

Schneider, U., Becker, A., Finger, P., Meyer-Christoffer, A., Rudolf, B., & Ziese, M. (2011). GPCC full data reanalysis version 6.0 at 0.5: Monthly land-surface precipitation from rain-gauges built on GTS-based and historic data; 2011.

Sellar, A. A., Jones, C. G., Mulcahy, J., Tang, Y., Yool, A., Wiltshire, A., et al. (2019). UKESM1: Description and evaluation of the UK Earth system model. *Journal of Advances in Modeling Earth Systems*, *11*, 4513–4558. https://doi.org/10.1029/2019MS001739

Sellar, A. A., Walton, J., Jones, C. G., Wood, R., Abraham, N. L., Andrejczuk, M., et al. (2020). Implementation of UK Earth system models for CMIP6. *Journal of Advances in Modeling Earth Systems*, *12*, e2019MS001946. https://doi.org/10.1029/2019MS001946

Sen Gupta, A., Jourdain, N. C., Brown, J. N., & Monselesan, D. (2013). Climate drift in CMIP5 models. *Journal of Climate*, *26*, 8597–8615. https://doi.org/10.1175/JCLI-D-12-00521.1

Seppala, A., Matthes, K., Randall, C., & Mironova, I. (2014). What is the solar influence on climate? Overview of activities during CAWSES-II. *Progress in Earth and Planetary Science*, *1*(1), 1–24.

Smeed, D. A., McCarthy, G., Cunningham, S. A., Frajka-Williams, E., Rayner, D., Johns, W. E., et al. (2013). Observed decline of the Atlantic Meridional Overturning Circulation 2004 to 2012. *Ocean Science Discussions*, *10*(5), 1619–1645. https://doi.org/10.5194/osd-10-1619-2013

Smith, D. M., Booth, B. B. B., Dunstone, N. J., Eade, R., Hermanson, L., Jones, G. S., et al. (2016). Role of volcanic and anthropogenic aerosols in the recent global surface warming slowdown. *Nature Climate Change*, *6*, 936–940. https://doi.org/10.1038/nclimate3058

Stenchikov, G., Hamilton, K., Stouffer, R. J., Robock, A., Ramaswamy, V., Santer, B., & Graf, H.-F. (2006). Arctic Oscillation response to volcanic eruptions in the IPCC AR4 climate models. *Journal of Geophysical Research*, *111*, D07107. https://doi.org/10.1029/2005JD006286

Stevenson, S., Fasullo, J. T., Otto-Bliesner, B. L., Tomas, R. A., & Gao, C. (2017). Role of eruption season in reconciling model and proxy responses to tropical volcanism. *Proceedings of the National Academy of Sciences*, *114*(8), 1822–1826. https://doi.org/10.1073/pnas.1612505114

Storkey, D., Blaker, A. T., Mathiot, P., Megann, A., Aksenov, Y., Blockley, E. W., et al. (2018). UK global ocean GO6 and GO7: A traceable hierarchy of model resolutions. *Geoscientific Model Development*, *11*(8), 3187–3213. https://doi.org/10.5194/gmd-11-3187-2018

Sutton, R. T., & Dong, B. (2012). Atlantic Ocean influence on a shift in European climate in the 1990s. *Nature Geoscience*, *5*(11), 788.

Sutton, R. T., & Hodson, D. L. R. (2005). Atlantic Ocean forcing of North American and European summer climate. *Science*, *309*(5731), 115–118.

Sutton, R. T., McCarthy, G. D., Robson, J., Sinha, B., Archibald, A. T., & Gray, L. J. (2018). Atlantic multidecadal variability and the U.K. ACSIS program. *Bulletin of the American Meteorological Society*, *99*(2), 415–425. https://doi.org/10.1175/BAMS-D-16-0266.1

Titchner, H. A., & Rayner, N. A. (2014). The Met Office Hadley Centre sea ice and sea surface temperature data set, version 2: 1. Sea ice concentrations. *Journal of Geophysical Research: Atmospheres*, *119*, 2864–2889. https://doi.org/10.1002/2013JD020316

Trenberth, K. E., & Shea, D. J. (2006). Atlantic hurricanes and natural variability in 2005. *Geophysical Research Letters*, *33*, L12704. https://doi.org/10.1029/2006GL026894

Undorf, S., Bollasina, M. A., Booth, B. B. B., & Hegerl, G. C. (2018). Contrasting the effects of the 1850-1975 increase in sulphate aerosols from North America and Europe on the Atlantic in the CESM. *Geophysical Research Letters*, *45*, 11,930–11,940. https://doi.org/10.1029/2018GL079970

Valcke, S., Craig, T., & Coquart, L. (2015). OASIS3-MCT user guide, OASIS3-MCT 3.0, technical report, TR/CMGC/15/38, CERFACS/CNRS SUC URA No. 1875, Toulouse, France.

Vaughan, D. G., Comiso, J., Allison, I., Carrasco, J., Kaser, G., Kwok, R., et al. (2013). Observations: Cryosphere. In Stocker, T. F., Qin, D., Plattner, G. K., Tignor, M., Allen, S. K., Boschung, J., & Nauels, A. (Eds.), *Climate change 2013: The physical science basis. Contribution of working group I to the fifth assessment report of the intergovernmental panel on climate change*: Cambridge University Press English (US).

Von Schuckmann, K., Palmer, M., Trenberth, K., Cazenave, A., Chambers, D., Champollion, N., et al. (2016). An imperative to monitor Earth's energy imbalance. *Nature Climate Change*, *6*(2), 138.

Walters, D., Baran, A. J., Boutle, I., Brooks, M., Earnshaw, P., Edwards, J., et al. (2019). The Met office unified model global atmosphere 7.0/7.1 and JULES global land 7.0 configurations. *Geoscientific Model Development*, *12*(5), 1909–1963. https://doi.org/10.5194/gmd-12-1909-2019

Watanabe, M., & Tatebe, H. (2019). Reconciling roles of sulphate aerosol forcing and internal variability in Atlantic multidecadal climate changes. *Climate Dynamics*, *53*(7), 4651–4665. https://doi.org/10.1007/s00382-019-04811-3

Webb, M. J., Andrews, T., Bodas-Salcedo, A., Bony, S., Bretherton, C. S., Chadwick, R., et al. (2017). The Cloud Feedback Model Intercomparison Project (CFMIP) contribution to CMIP6. *Geoscientific Model Development*, *10*(1), 359–384. https://doi.org/10.5194/gmd-10-359-2017

Weedon, G. P., Balsamo, G., Bellouin, N., Gomes, S., Best, M. J., & Viterbo, P. (2014). The WFDEI meteorological forcing data set: WATCH Forcing Data methodology applied to ERA-Interim reanalysis data. *Water Resources Research*, *50*, 7505–7514. https://doi.org/10.1002/2014WR015638

Williams, K. D., Copsey, D., Blockley, E. W., Bodas-Salcedo, A., Calvert, D., Comer, R., et al. (2017). The Met office global coupled model 3.0 and 3.1 (GC3.0 and GC3.1) configurations. *Journal of Advances in Modeling Earth Systems*, *10*, 357–380. https://doi.org/10.1002/2017MS001115

Wills, R. C. J., Armour, K. C., Battisti, D. S., & Hartmann, D. L. (2019). Ocean-atmosphere dynamical coupling fundamental to the Atlantic multidecadal oscillation. *Journal of Climate*, *32*(1), 251–272. https://doi.org/10.1175/JCLI-D-18-0269.1

Yan, X., Zhang, R., & Knutson, T. R. (2018). Underestimated AMOC variability and implications for AMV and predictability in CMIP models. *Geophysical Research Letters*, *45*, 4319–4328. https://doi.org/10.1029/2018GL077378

Zhang, R., & Delworth, T. L. (2006). Impact of Atlantic multidecadal oscillations on India/Sahel rainfall and Atlantic hurricanes. *Geophysical Research Letters*, *33*, L17712. https://doi.org/10.1029/2006GL026267

Zhang, R., Delworth, T. L., Sutton, R., Hodson, D. L. R., Dixon, K. W., Held, I. M., et al. (2013). Have aerosols caused the observed Atlantic multidecadal variability? *Journal of the Atmospheric Sciences*, *70*(4), 1135–1144. https://doi.org/10.1175/jas-d-12-0331.1

Zhang, J., & Rothrock, D. (2003). Modeling global sea ice with a thickness and enthalpy distribution model in generalized curvilinear coordinates. *Monthly Weather Review*, *131*(5), 845–861.

Zhang, Y., Wallace, J. M., & Battisti, D. S. (1997). ENSO-like interdecadal variability: 190093. *Journal of Climate*, *10*(5), 1004–1020. https://doi.org/10.1175/1520-0442(1997)010h1004:ELIVi2.0.CO;2