Supplement for "How well can we quantify when 1.5°C of global warming has been exceeded?"

Peter W. Thorne[1], John M. Nicklas[2], John J. Kennedy[3], Bruce Calvert[4], Baylor Fox-Kemper[2], Mark T. Richardson[5], Adrian Simmons[6], Ed Hawkins[7], Robert Rohde[8], Kathryn Cowtan[9], Nerilie J. Abram[10], Axel Andersson[11], Simon Noone[1], Phillipe Marbaix[12, 55], Nathan Lenssen[13] , Dirk Olonscheck[14], Tristram Walsh[15], Stephen Outten[16], Ingo Bethke[17], Bjorn H. Samset[18], Chris Smith[19,57], Anna Pirani[20], Jan Fuglestvedt[18], Lavanya Rajamani[21], Richard A. Betts[22], Elizabeth C. Kent[23], Blair Trewin[24], Colin Morice[22], Tim Osborn[25], Samantha N Burgess[6], Oliver Geden[26], Andrew Parnell[27], Piers M. Forster[28], Chris Hewitt[23,29], Zeke Hausfather[30], Valerie Masson-Delmotte[31], Jochem Marotzke[14], Nathan Gillett[32], Sonia I. Seneviratne[33], Gavin A. Schmidt[34], Duo Chan[35], Stefan Brönnimann[36], Andy Reisinger[37], Matthew Menne[38], Maisa Rojas Corradi[39], Christopher Kadow[40], Peter Huybers[41], David B. Stephenson[42], Emily Wallis[25], Joeri Rogelj[43], Andrew Schurer[44], Karen McKinnon[45], Panmao Zhai[46], Fatima Driouech[47], Wilfran Moufouma Okia[29], Saeed Vazifehkhah[29], Sophie Szopa[48], Christopher J. Merchant[49], Shoji Hirahara[50], Masayoshi Ishii[50,58], Francois A. Engelbrecht[51], Qingxiang Li[52], June-Yi Lee[53], Alex J. Cannon[54], C. Cassou[56], K. von Schuckmann[59], Amir H. Delju[29], Ellie Murtagh[1]

[1] ICARUS Climate Research Centre, Maynooth University, Maynooth, Co. Kildare, Ireland

[2] Department of Earth, Environmental, and Planetary Sciences, Brown University, Providence, RI, USA

[3] Independent researcher, France

[4] Independent Researcher, Ottawa, Canada

[5] Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA, USA

[6] European Centre for Medium-Range Weather Forecasts, Reading, UK.

[7] National Centre for Atmospheric Science, Department of Meteorology, University of Reading, Reading, UK.

[8] Berkeley Earth, Berkeley, CA, USA

[9] Department of Chemistry, University of York, York, UK

[10] ARC Centre of Excellence for the Weather of the 21st Century, Research School of Earth Sciences, The Australian National University, Canberra ACT 2601, Australia.

[11] German Meteorological Service, Hamburg, Germany

[12] Division of Thermodynamics and Fluid Dynamics, Université catholique de Louvain, Louvain-la-Neuve, Belgium

[13] Colorado School of Mines, Golden Colorado, USA & NSF National Center for Atmospheric Research, Boulder, Colorado, USA

[14] Max Planck Institute for Meteorology, Hamburg, Germany

[15] Environmental Change Institute, University of Oxford, Oxford, UK

[16] Nansen Environmental and Remote Sensing Center, Bjerknes Centre for Climate Research, Bergen, Norway https://orcid.org/0000-0002-4883-611X

[17] University of Bergen, Bjerknes Centre for Climate Research, Bergen, Norway

[18] CICERO Center for International Climate Research, Oslo, Norway

[19] Department of Water and Climate, Vrije Universiteit Brussel, Brussels, Belgium

[20] Euro-Mediterranean Centre on Climate Change (CMCC Foundation), Venice, Italy

[21] Faculty of Law, University of Oxford

[22] Met Office Hadley Centre, Fitzroy Road, Exeter, UK

[23] National Oceanography Centre, Southampton, UK

[24] Bureau of Meteorology, Melbourne, Australia

[25] Climatic Research Unit, School of Environmental Sciences, University of East Anglia, Norwich, UK

[26] German Institute for International and Security Affairs, Berlin, Germany

[27] School of Mathematics and Statistics, University College Dublin, Dublin, Ireland

[28] Priestley Centre for Climate Futures, University of Leeds, Leeds, UK

[29] WMO Secretariat, World Meteorological Organization, Geneva, Switzerland

[30] Stripe Inc., South San Francisco, CA, USA and Berkeley Earth, Berkeley, CA, USA

[31] Laboratoire des Sciences du Climat et de l'Environnement (UMR 8212 CEA-CNRS-UVSQ), Institut Pierre Simon Laplace, Université Paris-Saclay, France

[32] Canadian Centre for Climate Modelling and Analysis, Environment and Climate Change Canada, Victoria, BC, Canada.

[33] Institute for Atmospheric and Climate Science, ETH Zurich, Zurich, Switzerland

[34] NASA Goddard Institute for Space Studies, New York, NY, USA

[35] School of Ocean and Earth Science, University of Southampton, UK

[36] Institute of Geography and Oeschger Centre for Climate Change Research, University of Bern, Switzerland

[37] Institute for Climate, Energy and Disaster Solutions, Australian National University, Acton ACT, Australia

[38] NOAA's National Centers for Environmental Information, 151 Patton Avenue, Asheville, NC, USA

[39] Department of Geophysics, Faculty of Engineering, University of Chile, Santiago, Chile.

[40] German Climate Computing Center (DKRZ), Hamburg, Germany

[41] Department of Earth and Planetary Sciences, Harvard University, Cambridge, MA, USA

[42] Department of Mathematics and Statistics, University of Exeter, Exeter, UK

[43] Grantham Institute - Climate Change and the Environment, and Centre for Environmental Policy, Imperial College, London, UK

[44] School of GeoSciences, University of Edinburgh, UK

[45] Department of Statistics and Data Science, Department of Atmospheric and Oceanic Sciences, Institute of the Environment and Sustainability, University of California, Los Angeles

[46] State Key Laboratory of Severe Weather Meteorological Science and technology, Chinese Academy of Meteorological Sciences, Beijing, China

[47] University Mohammed VI Polytechnic, Ben Guerir, Morocco

[48] Laboratoire des Sciences du Climat et de l'Environnement (UMR 8212 CEA-CNRS-UVSQ), Institut Pierre Simon Laplace, Université Paris-Saclay, France

[49] National Centre for Earth Observation, University of Reading, Reading, UK

[50] Meteorological Research Institute, Tsukuba, Japan

[51] Global Change Institute, University of the Witwatersrand, Johannesburg, South Africa

[52] Sun Yat-Sen University, Guangzhou/Zhuhai, China

[53] Research Center for Climate Sciences, Pusan National University and Center for Climate Physics, Institute for Basic Science, Busan, Republic of Korea
[54] Climate Research Division, Environment and Climate Change Canada, Victoria, BC, Canada.
[55] Department of Geography, University of Liège, Belgium
[56]LMD-IPSL, CNRS, Ecole Normale Superieure, PSL Research University, Paris, 75005, France
[57] Energy, Climate and Environment Program, International Institute for Applied Systems Analysis, Laxenburg, Austria
[58] Geoenvironmental Sciences, University of Tsukuba, Tsukuba, Japan
[59] Mercator Ocean international, Toulouse, France

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

This Supplement contains supplementary information in support of the main paper. Supplementary material is provided principally for Sections 4.6 and 5 where further details that were not possible to fit within the main analysis are given.

1 Commonly used acronyms

Table S1 lists acronyms used frequently in the main paper and their definitions.

Acronym	**Acronym longform / description and key references where necessary**
CGWL	Current Global Warming Level (specifically Betts et al. 2023 method)
ENSO	El Niño Southern Oscillation
EOF	Empirical Orthogonal Function
EOT	Empirical Orthogonal Teleconnection
ESM	Earth System Model
FaIR	Finite Amplitude Impulse Response simple climate model (Smith et al., 2017, 2024 method)
GMST	Global Mean Surface Temperature
GSAT	Global Surface Air Temperature

GWI	Global Warming Index (Haustein et al., 2017 & Forster et al. 2024 method)
GWL	Global Warming Level
ICV	Internal Climate Variability
ICJ	International Court of Justice
IMO-WMO (sometimes just WMO)	International Meteorological Organisation / World Meteorological Organization
IPCC	Intergovernmental Panel on Climate Change
KCC	Kriging for Climate Change (Qasmi and Ribes 2022 method)
LSAT	Land Surface Air Temperatures
LTTG	Long-Term Temperature Goal
MAT	Marine Air Temperatures
NMAT	Night Marine Air Temperatures
OLS	Ordinary Least Squares regression
PA	Paris Agreement
PAGES	Past Global Changes
PDF	Probability Density Function
RFC	Reasons for Concern
ROF	Regularised Optimal Fingerprinting (Gillett et al., 2021 method)
SED	Structured Expert Dialogue
SLCF	Short Lived Climate Forcers
SMILE	Single Model Initialised Large Ensemble
SST	Sea Surface Temperature
UNFCCC	United Nations Framework Convention on Climate Change
VEI	Volcanic Explosivity Index

Table S1. List of acronyms commonly used in the paper

2 Supplementary Information for Section 4.6

2.1 Summary of methods considered in Section 4.6

The methods used throughout Section 4.6 are outlined in Table S2 with back references to the sections in Section 4 in which they are either discussed or the method class is outlined. While very many methods have been considered this set is not entirely comprehensive. Some published methods we were unable to contact the authors or to get working. A further subset of methods were also unable to be modified to be applied to the synthetic future cases detailed in Section 4.6.2.

Code to be able to run all methods run in this analysis, except for GWI, are made available at https://github.com/jnickla1/Thorne_15. GWI approach information can be found at https://github.com/tristramwalsh/global-warming-index

Method short name used in figures in main text	Description	Reference	Limited application to or $ if included in its own figure	Described in main text section
cent20y	Average of 19- and 21--yr centred		$(Fig 14)	4.3/1_Run_Means
cent21y	21-yr centred			4.3/1_Run_Means
cent30y	Average of 29- and 31--yr centred		$(Fig 14)	4.3/1_Run_Means
lag5y	5-yr lagging			4.3/1_Run_Means
lag10y	10-yr lagging	AR6	$(Fig 14)	4.3/1_Run_Means
OLS_refit	OLS refitting	AR5	$(Fig 14)	4.3/2_LT_Fits
OLS_AR5all	OLS AR5 all	AR5		4.3/2_LT_Fits
OLS_AR5split	OLS AR5 split			4.3/2_LT_Fits
TheilSen_h7075	Theil Sen slope after hinge fit 1975	Duan et al., 2021		4.3/2_LT_Fits
OLS_hinge75	OLS Hinge fit 1975 (refitting mt+b)	Livezey et. al. 2007		4.3/2_LT_Fits
hinge75meet	Hinge fit meet (refitting mt, b fixed)	(to make lines meet)	$(Fig 14)	4.3/2_LT_Fits
quartic	quartic polynomial	Hawkins and Sutton 2009		4.3/2_LT_Fits
Bayes_seq_CP	Bayesian change-point	Yu and Ruggieri, 2019		4.3/2_LT_Fits
offset11y	11-yr offset	Trewin (2022)	$(Fig 14)	4.3/3_ST_Fits
etrend15y	15-yr trend endpoint	SR15 Allen et al., 2018	$(Fig 14)	4.3/3_ST_Fits
etrend30y	30-year trend endpoint			4.3/3_ST_Fits
etrend30y_C3S	End of 30-year trend C3S			4.3/3_ST_Fits
min_month_proj	Projection into the future from the minimum monthly temperature	Cannon, 2025 Bevacqua et al., 2025	obs	4.3/3_ST_Fits

	observed over the past year			
lowess1dt10wnc lowess1dt20wnc lowess1dt26wnc lowess1dt30wnc lowess1dt36wnc	LOWESS linear, tricube kernel (width 10, 20, or 30), standard error not corrected	Clarke & Richardson (2021)		4.3/3_ST_Fits
lowess2dt20wnc	LOWESS quadratic, tricube kernel, width 20, standard error not corrected			4.3/3_ST_Fits
lowess1dg20wnc	LOWESS linear, gaussian kernel, width 20, standard error not corrected			4.3/3_ST_Fits
lowess1dt20wAR	LOWESS linear, tricube kernel (width 20), standard error corrected via AR(1) coefficient	Clarke & Richardson (2021) recommended (eq 5, 6, 11)		4.3/3_ST_Fits
lowess1dt20wARMA	LOWESS linear, tricube kernel (width 20), standard error corrected via ARMA parameters using MLE	Clarke & Richardson 2021, referencing Hausfather 2017		4.3/3_ST_Fits
butterworth	Butterworth	Mann (2008)	$(Fig 14)	4.3/2_LT_Fits
	Empirical Mode Decomposition	Wu 2011	none	4.3/3_ST_Fits
opt_clim_norm	Optimal climate normal	Livezey et. al. 2007		4.3/3_ST_Fits
cubic_spline	Cubic spline	Vissier 2018		4.3/4_GAM_AR1
GAM_AR1	GAM AR1 residuals	AR5, box 2.2	$(Fig 14)	4.3/4_GAM_AR1
GAM_AR0	GAM AR0 (standard)			4.3/2_LT_Fits
Kalman_RW Kalman_RW_ocn Kalman_EM_linRW	Kalman: Std Random Walk on GMST alone, on GMST and GSST, and using expectation maximization (EM) on the parameters of GMST & GSST	Shumway and Stoffer (2016)		4.3/5_Kalman
Kal_flexLin Kal_flexLin_ocn	Kalman: Integrated Rand Walk (on GMST alone, GMST & GSST)	Visser 2018	$(Fig 14)	4.3/5_Kalman
removeMEI_cons	Remove MEI	Foster and Rahmstorf 2011	$(Fig 14)	4.3/6_Remove_IV
removeMEI_volc_cons removeMEI_volc_refit	Remove MEI, volcanic AOD, solar. Cons (constant) or refit refers to	Foster and Rahmstorf 2011		4.3/6_Remove_IV

	whether the coefficients of each of these linearized components are recalculated at each timestep or held constant to the values provided in the paper.			
	Atlantic and Pacific modes of variability	Wu et al. 2019	none	4.3/6_Remove_IV
lfca_SST lfca_hadcrut	Low-frequency component analysis (LFCA) is a method that transforms the leading empirical orthogonal functions (EOFs). Analyze either SST or HadCRUT.	Wills et al. 2018; Wills et al. 2020		4.3/6_Remove_IV
	Spatial EOFs of TS	Chen and Tung, 2018	none	4.3/6_Remove_IV
	Reg. Linear Models of Sea level pressure	Sippel et al. 2019	none	4.3/6_Remove_IV
removeGreensfx	Greens Functions on TS	Samset et al. 2023. Thanks to Bjørn Samset for running these computations.		4.3/6_Remove_IV
CGWL10y_pUKCP CGWL10y_sUKCP CGWL10y_sfUKCP	UKCIP18 RCP4.5 CGWL. "p" refers to the entire ensemble probabilistic prediction, "s" refers to subsampling this ensemble to match those closest to last year's temperature, and "sf" refers to subsampling to match both last year's temperature and the past decadal average.	Betts et. al. (2023)	$(Fig 17)	4.5_EarthModel_CGWL
	HadGEM3 CGWL	Betts et. al. (2023)	none	4.5_EarthModel_CGWL
CGWL10y_forec CGWL10y_for_halfU	WMO Lead Centre for Annual-to-Decadal Climate Prediction (forecast coupled models) 5 CGWL. We also halved the uncertainty (halfU) because the ensemble was more dispersed than the error from the 20-year running mean.	Betts et. al. (2023), also thanks to Leon Hermanson	$(Fig 17)	4.5_EarthModel_CGWL

CGWL_10y_IPCC	IPCC Assessed warming trend. CGWL			4.5_EarthModel_CGWL
cons_hArrh_CO2forc OLS_refit_CO2forc	linear CO2-temperature anomaly analysis: constant coefficients (uncertainty from half Arrhenius's 5.5°C/doubling estimate) or OLS re-fit every year	Jarvis & Forster 2024		4.4/1_Linear
FaIR_all FaIR_all_unB FaIR_antho FaIR_antho_unB FaIR_nonat FaIR_nonat_unB FaIR_comb_unB	Effective Radiative Forcings via Finite-amplitude Impulse Response Model: FaIR with pre-calibration. Underlying MCMC on sensitivities and key climate parameters. Output is then produced from forcings (all, nonat) or these are combined together: (anthro = all - nonat, or comb = 2/3 * nonat + 1/3 * all). unB refers to post-hoc removal of both the overall and linear-trending bias due to a probability underflow issue in future tests.	Milllar et al., 2017 Smith et al., 2017 Smith et al., 2024 AR6 Chapter 7	$(Fig 15)	4.3/2_ERF_FaIR
EBMKF_ta	Nonlinear EBM within Kalman Filter (volcanos time-averaged, just literature parameters)	Nicklas et al. (2025)		4.3/3_Kalman
EBMKF_ta2	Nonlinear EBM (cloud sensitivity increased) within Kalman Filter			4.3/3_Kalman
EBMKF_ta4	Nonlinear EBM within Kalman Filter, additional top of atmosphere net radiation and unknown energy flux to compensate for ESM discrepancies		$(Fig 15)	4.3/3_Kalman
	3-layer linear heat Kalman Filter	Cummins (2020)	none, but does form a core part of FaIR	4.3/3_Kalman
GWI_tot GWI_tot_orig GWI_anthro GWI_anthro_orig	GWI: Attributed global warming using Global Warming Index multifingerprinting method: extracted either the total forced warming signal or the anthropogenic component as a	Otto et al. 2015, Haustein et al., 2017, Forster et al. 2024	$(Fig 15)	4.3/4_Human_Induced

	simple timeseries of annual mean values. Tot_orig and anthro_orig are the original computations performed for the Indicators of Global Climate Change (IGCC) 2023.			
GWI_tot_AR6 GWI_tot_CGWL GWI_tot_SR15 GWI_anthro_AR6 GWI_antrho_CGWL GWI_anthro_SR15	GWI: Attributed global warming: annual mean timeseries (see above) combined with additional multi-decadal averaging methods, namely: (i) "AR6": the formal IPCC AR6 WGI definition for attributed global warming (and its annual updates in the IGCC), which used the GWI as one of three methods. Definition: 10-year lagged average (ii) "SR15": the formal IPCC SR1.5 definition of the level of global warming (and its annual updates in the IGCC), which specifically used the GWI. Definition: 30-year average of human-induced warming centred by extrapolating the most recent 15 year trend into the future; technically only the "anthro" component is the strict definition used by SR1.5, but the other warming components (e.g. "tot") are also calculated here in the same way for comparison. (iii) "CGWL": the 20-year mean centred using the constrained future projections of the GWI timeseries, providing an analogous approach to the Betts et al., CGWL, except for for attributable warming instead of realised warming (as in the Betts et al. 2024 method).	Tristram Walsh, work performed specifically for this paper		4.3/4_Human_Induced
KCC_all KCC_human	Kriging for Climate Change (all and human-attributable components, as above)	Qasmi and Ribes 2022	obs $(Fig 15)	4.3/4_Human_Induced

eROF_anthro	Regularized Optimal Fingerprinting	Gillett et al. 2021	obs, note**, $(Fig 15)	4.3/4_Human_Induced
	Neural Network (CMIP trained)	Bone et al., 2023a	none	Section 7
	UNet (preprint)	Bone et al., 2023b https://essopenarchive.org/users/653004/articles/660203-separation-of-internal-and-forced-variability-of-climate-using-a-u-net	none	Section 7
	Artificial neural networks (ANNs)	Diffenbaugh & Barnes, 2023	none	Section 7

Table S2. Summary of the methods considered for inclusion throughout Section 4.6. "None" methods in the 4th column were not evaluated in any tests. Note ** eROF was technically not a current method since future years contribute to fitted parameters, but could be adapted to a strictly current test. eROF, KCC, and GWI can't be evaluated in future volcano tests since that would require thousands of CMIP simulations.

Min_month_proj modifications:

The approach based on Cannon (2025) extrapolates from the minimum monthly temperature observed within recent 12-month periods to estimate when long-term warming thresholds will be crossed.

The method operates by fitting 15-year linear trends to monthly temperature data, then projecting forward from the coolest month in each trailing year. The projection incorporates three key parameters from Cannon (2025): a 33-month offset representing the typical lag between short-term temperature extremes and long-term threshold crossing, plus uncertainty terms accounting for the 90% confidence range in this timing (-28 to +76 months) and observational uncertainty in the temperature record itself.

For future climate scenarios using ESM1-2-LR or NorESM models, the code constructs continuous temperature records by concatenating historical simulations with future projections, carefully matching baselines between observed and simulated periods. A critical preprocessing step removes the seasonal cycle from these merged datasets to isolate the underlying warming trend. The method then tracks a 4-year rolling maximum of annual minimum temperatures, propagating both the central projection and combined uncertainties from trend estimation, timing variability, and measurement error through the calculation.

2.2 Further analysis of historical exceedances (Section 4.6.1)

The main paper in Figure 20 illustrated the degree of agreement of candidate methods with a retrospectively calculated 20-year average for the exceedance of 1°C warming. Here we repeat the analysis for the exceedance of 0.5°C warming. This exceedance occurred around 1985 in HadCRUT5 (Figure S1) and similarly in the mid-1980s to early 1990s across all products (not shown). Owing to the presence of the Pinatubo and el Chichon eruptions relatively close to this date in several datasets the annual mean temperature exceeds 0.5°C then dips below before exceeding on a more permanent basis. Here we benchmark against the only time it is exceeded in the 20-year centered mean[1], which has a sufficiently wide averaging window that it has increased monotonically since 1965 (~0.25°C of warming). Several of the methods assessed here that include volcanic influences (from Section 4.4) similarly crossed 0.5°C up to three times (not shown), and the 20-year centered mean may fall below future thresholds (eg. 1.5°C) if a sufficiently large volcanic eruption occurs (see section 4.6.2).

The violin plots in Figure S1 and Figure 20 were created by fitting a cubic spline to the inverse-gaussian of the probability of exceedance data. This method is equivalent to adding a pre-function to correct the q-q plot. Once this smooth cubic spline fit is found, manipulation such as remapping, plotting, or conversion between the pdf and cumulative distribution function is straightforward.

Overall results for 0.5°C exceedance are similar to those for 1°C exceedance (Figure S1). Methods that performed particularly poorly at 1°C also tend to perform poorly for exceedance of 0.5°C. Notably OLS linear fit, while still constituting a statistical 'miss', is somewhat closer. This aligns with IPCC AR5 (Hartmann et al., 2013) where the 10-year lagging mean relative to 1850-1900 and an OLS trend fit from 1880 almost exactly matched for the period ending in 2012.

Given this detailed examination of 0.5 and 1°C exceedances, a metric for continued consideration of each method based on the RMS error (Table S3) was utilized for simplicity. Examination of this table reveals that essentially the same set of estimates would be used if log-likelihood were the deciding metric.

method_name	method_class	RMS	log-likel	#q<.5	bias	cross_yr0.5	cross_yr1.0	#yrs
FaIR_nonat_unB	44/2_ERF_FaIR	0.02475	2.313	0	0.00001	1984.9	2010.3	95
FaIR_nonat	44/2_ERF_FaIR	0.02475	2.305	0	0.00001	1984.8	2010.2	95

[1] As the midpoint of a true 20-year centered mean would fall on a half-year time point, we define within this paper a 20-year centered mean to be the average of a 19-year and 21-year centered mean to compare to the variety of other methods which report warming at whole integer years.

GWI_tot_CGWL	44/4_Human_Induced	0.02632	2.222	0	-0.00001	1985.6	2009.5	75
FaIR_comb_unB	44/2_ERF_FaIR	0.02706	2.189	2	-0.00165	1985.4	2009.3	91
FaIR_anthro_unB	44/2_ERF_FaIR	0.02728	2.178	9	0.00019	1983.6	2010.5	95
EBMKF_ta2	44/3_Kalman	0.02769	2.146	0	0.00005	1984.7	2010.3	174
FaIR_anthro	44/2_ERF_FaIR	0.02770	2.166	0	-0.00481	1983.8	2010.7	95
EBMKF_ta	44/3_Kalman	0.02942	2.075	0	0.00694	1983.7	2009.1	174
EBMKF_ta4	44/3_Kalman	0.03024	2.056	0	-0.00793	1985.7	2010.7	174
CGWL10y_for_halfU	45_EarthModel_CGWL	0.03330	1.988	4	-0.01946	1987.4	2009.8	64
CGWL10y_forec	45_EarthModel_CGWL	0.03330	1.957	0	-0.01946	1987.5	2010.0	64
GWI_anthro_AR6	44/4_Human_Induced	0.03417	1.956	0	-0.00003	1986.2	2012.7	75
TheilSen_h7075	43/2_LT_Fits	0.03657	2.048	0	0.00404	1980.7	2012.4	90
hinge75meet	43/2_LT_Fits	0.03722	1.893	1	0.00492	1980.2	2010.4	90
CGWL10y_pUKCP	45_EarthModel_CGWL	0.03957	-4.105	64	0.00785	1987.0	2009.9	165
GWI_anthro_CGWL	44/4_Human_Induced	0.04011	1.825	8	0.00004	1983.5	2010.9	75
GWI_anthro	44/4_Human_Induced	0.04107	1.814	8	-0.00001	1984.1	2011.5	75
CGWL10y_sfUKCP	45_EarthModel_CGWL	0.04113	1.742	7	0.00143	1985.7	2009.6	155
CGWL10y_sUKCP	45_EarthModel_CGWL	0.04271	1.746	4	0.00016	1985.5	2010.1	155
OLS_hinge75	43/2_LT_Fits	0.04302	1.982	2	0.00743	1979.5	2011.0	90
GWI_anthro_SR15	44/4_Human_Induced	0.04318	1.745	5	0.00001	1985.5	2011.4	75
CGWL_10y_IPCC	45_EarthModel_CGWL	0.04380	1.677	0	-0.00021	1983.2	2009.7	166
Kal_flexLin_ocn	43/5_Kalman	0.04424	1.687	0	-0.00408	1986.9	2011.8	175

Kal_flexLin	43/5_Kalman	0.04813	1.610	0	-0.00939	1986.5	2012.8	175
removeGreensfx	43/6_Remove_IV	0.04933	1.590	0	-0.00170	1980.6	2009.9	174
Kalman_RW_ocn	43/5_Kalman	0.04969	1.577	0	-0.02171	1987.4	2014.2	175
lowess1dt30wnc	43/3_ST_Fits	0.05053	1.463	2	-0.01072	1985.5	2012.0	172
lowess1dt36wnc	43/3_ST_Fits	0.05104	1.429	0	-0.01319	1986.6	2011.7	172
lowess1dt26wnc	43/3_ST_Fits	0.05238	1.431	1	-0.00924	1982.8	2012.6	172
Kalman_RW	43/5_Kalman	0.05277	1.518	0	-0.02264	1987.3	2014.5	175
removeMEI_volc_cons	43/6_Remove_IV	0.05399	1.488	0	-0.01261	1980.5	2008.7	53
butterworth	43/2_LT_Fits	0.05587	1.237	3	-0.03451	1982.8	2013.8	76
etrend30y	43/3_ST_Fits	0.05589	1.438	0	-0.01114	1987.5	2011.1	146
lag5y	43/1_Run_Means	0.05629	1.344	0	-0.01802	1982.5	2014.9	171
lowess1dt20wAR	43/3_ST_Fits	0.05745	1.273	2	-0.00751	1982.1	2013.9	172
lowess1dt20wnc	43/3_ST_Fits	0.05745	1.304	2	-0.00751	1982.1	2013.9	172
GAM_AR1	43/4_GAM_AR1	0.05810	1.241	17	-0.01201	1986.4	2012.6	146
lowess1dt20wARMA	43/3_ST_Fits	0.05839	1.318	2	-0.00408	1982.1	2013.9	150
etrend15y	43/3_ST_Fits	0.05884	1.373	2	-0.00613	1982.3	2013.0	161
GWI_tot_SR15	44/4_Human_Induced	0.05945	1.378	0	0.00005	1986.6	2006.1	75
lag10y	43/1_Run_Means	0.06087	1.223	0	-0.03263	1987.6	2015.8	166
GWI_tot	44/4_Human_Induced	0.06155	1.394	6	0.00002	1986.5	2009.9	75
GWI_tot_AR6	44/4_Human_Induced	0.06183	1.345	2	0.00005	1987.0	2011.9	75
FaIR_all_unB	44/2_ERF_FaIR	0.06447	1.322	2	0.00006	1987.0	2006.8	95

opt_clim_norm	43/3_ST_Fits	0.06468	0.984	12	-0.02112	1980.8	2013.9	175
etrend30y_3CS	43/3_ST_Fits	0.06474	1.264	0	-0.02047	1987.5	2015.5	146
lowess1dg20wnc	43/3_ST_Fits	0.06490	1.169	0	-0.02386	1989.2	2012.8	172
GAM_AR0	43/4_GAM_AR1	0.06521	1.323	12	-0.00611	1982.9	2012.9	146
FaIR_all	44/2_ERF_FaIR	0.06814	1.271	2	0.02206	1985.7	2005.6	95
lfca_TAS	43/6_Remove_IV	0.06831	1.266	0	-0.04648	1986.6	2013.8	48
GWI_tot_orig	44/4_Human_Induced	0.06831	1.245	7	0.00001	1984.8	2008.4	74
lowess2dt30wnc	43/3_ST_Fits	0.06876	1.172	0	-0.00479	1980.5	2014.2	172
KCC_human	44/4_Human_Induced	0.07200	1.021	20	-0.00002	1983.9	2009.7	124
eROF_tot	44/4_Human_Induced	0.07383	1.031	39	-0.00004	1987.2	2008.0	175
lowess1dt10wnc	43/3_ST_Fits	0.07421	0.988	0	-0.00580	1980.0	2014.0	172
lowess2dt20wnc	43/3_ST_Fits	0.07627	1.003	1	-0.00555	1980.0	2014.1	171
Bayes_seq_CP	43/2_LT_Fits	0.07638	1.100	0	-0.03116	1989.3	2012.7	126
eROF_anthro	44/4_Human_Induced	0.07794	0.915	25	-0.00004	1987.1	2008.0	175
cubic_spline	43/4_GAM_AR1	0.07935	1.114	0	-0.00466	1980.0	2015.1	145
GWI_anthro_orig	44/4_Human_Induced	0.08049	1.119	0	0.00002	1982.4	2008.7	74
removeMEI_cons	43/6_Remove_IV	0.08188	0.630	4	0.00524	1979.9	2008.7	152
OLS_refit_CO2forc	44/1_Linear	0.08642	-3.842	59	-0.01811	1985.5	2012.2	175
offset11y	43/1_Run_Means	0.09200	0.999	62	0.07450	1982.6	2009.9	165
min_month_proj	43/3_ST_Fits	0.09458	1.026	1	0.02113	1979.7	2009.5	156
KCC_all	44/4_Human_Induced	0.09746	0.968	4	0.00003	1988.0	2011.0	124

raw1y	43/1_Run_Means	0.09798	0.898	1	-0.00538	1979.5	2009.5	175
lfca_SST	43/6_Remove_IV	0.09800	0.926	0	-0.05228	1989.2	2014.7	77
quartic	43/2_LT_Fits	0.09906	0.992	0	-0.06507	1989.5	2009.5	75
cons_hArrh_CO2forc	44/1_Linear	0.10426	-85.493	89	0.07524	1978.4	2009.1	175
removeMEI_volc_refit	43/6_Remove_IV	0.11195	0.804	0	0.00909	1985.5	2006.6	45
OLS_refit	43/2_LT_Fits	0.13422	0.561	0	-0.03125	1989.7	2022.5	75

Table S3: Detailed results for every method historical comparison to the 20-yr centred running mean within HadCRUT5. Note that the uncertainty has been pre-scaled to optimize the log-likelihood from 1925-2024. This means that the average log-likelihood listed in the 4th column above could be further optimized for the entire window for methods that extend prior to 1925. Given the present pre-scaling, the average log-likelihood is substantially better over the 1925-2024 window (>0.3 change) than over this entire window as listed above for only a few methods: CGWL10y_pUKCP, OLS_refit_CO2forc, cons_hArrh_CO2forc. The red shading indicates the methods that do not pass the RMSE threshold condition.

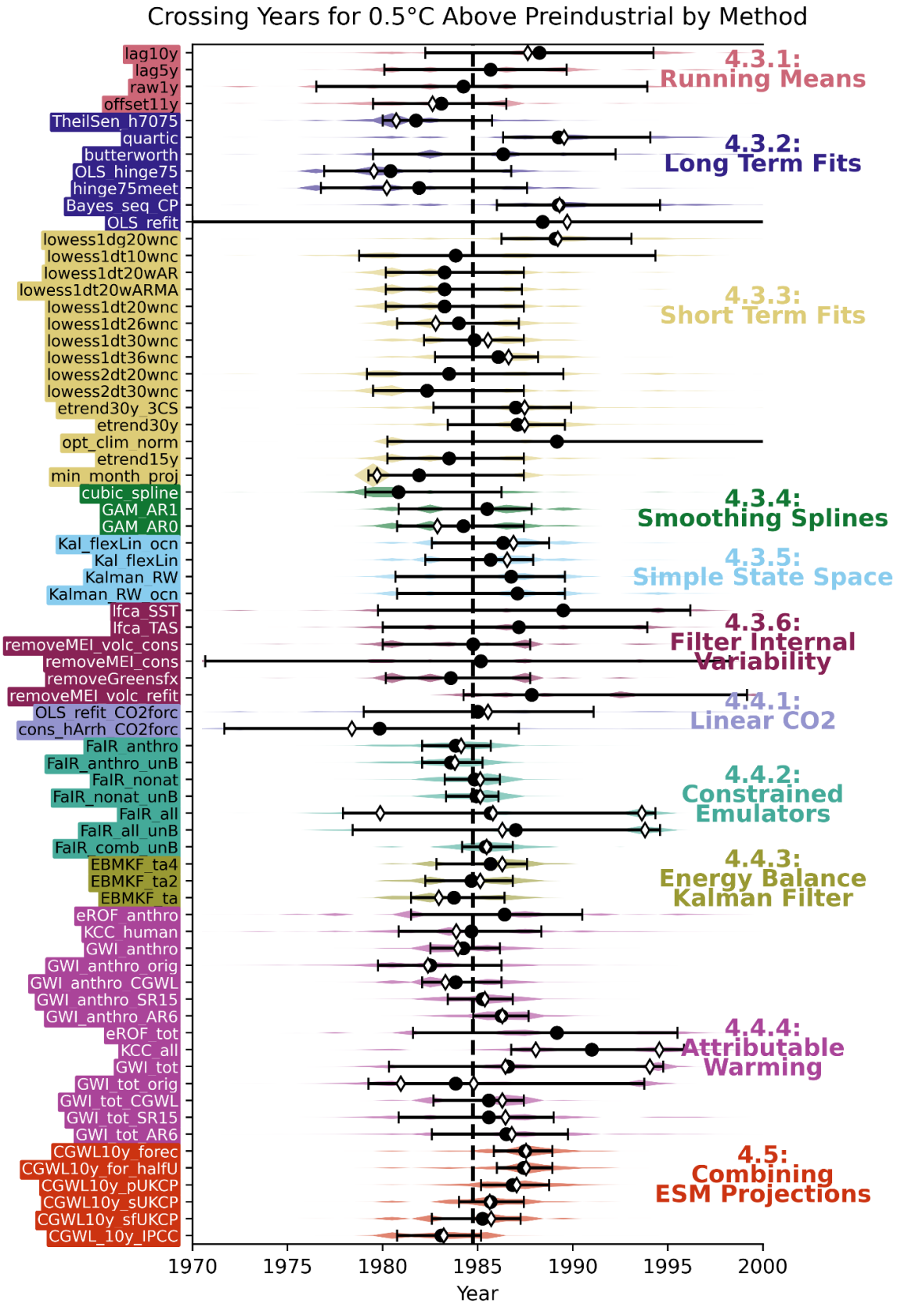
Crossing Years for 0.5°C Above Preindustrial by Method
lag10y
lag5y
raw1y
offset11y
TheilSen_h7075
quartic
butterworth
OLS_hinge75
hinge75meet
Bayes_seq_CP
OLS_refit
lowess1dg20wnc
lowess1dt10wnc
lowess1dt20wAR
lowess1dt20wARMA
lowess1dt20wnc
lowess1dt26wnc
lowess1dt30wnc
lowess1dt36wnc
lowess2dt20wnc
lowess2dt30wnc
etrend30y_3CS
etrend30y
opt_clim_norm
etrend15y
min_month_proj
cubic_spline
GAM_AR1
GAM_AR0
Kal_flexLin_ocn
Kal_flexLin
Kalman_RW
Kalman_RW_ocn
lfca_SST
lfca_TAS
removeMEI_volc_cons
removeMEI_cons
removeGreensfx
removeMEI_volc_refit
OLS_refit_CO2forc
cons_hArrh_CO2forc
FaIR_anthro
FaIR_anthro_unB
FaIR_nonat
FaIR_nonat_unB
FaIR_all
FaIR_all_unB
FaIR_comb_unB
EBMKF_ta4
EBMKF_ta2
EBMKF_ta
eROF_anthro
KCC_human
GWI_anthro
GWI_anthro_orig
GWI_anthro_CGWL
GWI_anthro_SR15
GWI_anthro_AR6
eROF_tot
KCC_all
GWI_tot
GWI_tot_orig
GWI_tot_CGWL
GWI_tot_SR15
GWI_tot_AR6
CGWL10y_forec
CGWL10y_for_halfU
CGWL10y_pUKCP
CGWL10y_sUKCP
CGWL10y_sfUKCP
CGWL_10y_IPCC
4.3.1:
Running Means
4.3.2:
Long Term Fits
4.3.3:
Short Term Fits
4.3.4:
Smoothing Splines
4.3.5:
Simple State Space
4.3.6:
Filter Internal
Variability
4.4.1:
Linear CO2
4.4.2:
Constrained
Emulators
4.4.3:
Energy Balance
Kalman Filter
4.4.4:
Attributable
Warming
4.5:
Combining
ESM Projections
1970
1975
1980
1985
1990
1995
2000
Year

Figure S1. As in Figure 20 in the main paper, but for exceedance of 0.5°C. See figure caption in main text for further explanations.

Given that no single approach is likely to be optimal in all circumstances, it is instructive to consider how the entire family of assessed approaches fared in determining the time of exceedance of both 0.5°C and 1°C (Figure S2). In both cases the methods as a whole bracket the actual time of exceedance with a majority of approaches within +/-3 years for 0.5°C and +/-2 years for 1°C. For exceedance of 1°C most methods determine the crossing to have occurred later than the actual retrospective 20-year average occurrence. Ignoring the extreme outlier arising from OLS fit in the mid-2020s for 1°C all methods for both exceedances are within +/-6 years.

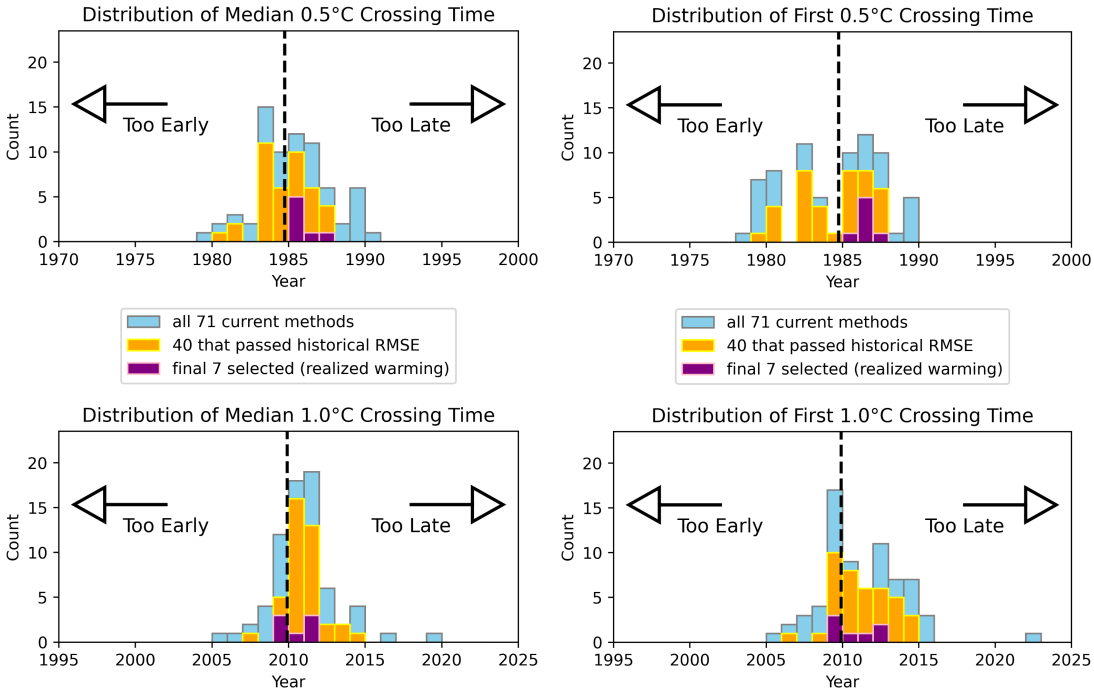

Figure S2. Overall distribution of the timing of estimates of crossing times relative to the 20-year lagged mean crossing times (shown by a dashed vertical line) in HadCRUT5. Different colours denote retention at different stages of the testing steps outlined in Section 4.6 in the main text.

2.3 Further analysis of possible future behaviour (Section 4.6.2)

2.3.1 Selection of methods able to be run against the synthetic test cases

It was not possible (or in some cases, not worthwhile) to modify all techniques to run on the future test cases that rely solely on model simulations and associated forcings. Those methods which were unable to be employed (see Table S2) and the reasons are as follows:

- lfca_TAS and lfca_SST (4.3.6: Filtering Internal Variability) required significant computational resources to analyze each ensemble member in the 270-member future scenarios suite. We didn't make this investment as this method was very poorly-performing in the historical evaluation (worse than the raw 1-year signal).
- GWI_tot_orig and GWI_tot_anthro (4.4.6: Human Induced): These were the original outputs from old code as applied to the historical HadCRUT5 (as published in the 2023 IGCC). The code was since updated by Tristram Walsh to perform substantially better, and we did not see the point of running old code on all the future ensemble members.
- eROF_tot and eROF_antrho. Had difficulty separating the ROF method from the EsmValTool package, so historical runs are technically not current but retrospective. Furthermore, this method (as applied naively) would require a CMIP ensemble to be run on each of the future volcanic forcings from the NorESM Voc ensemble, much like the hist-volc ensemble runs.
- KCC_all and KCC_human: much like ROF, would require a CMIP ensemble to be run on each of the future volcanic forcings from the NorESM Voc ensemble. Was beyond our available resources / time constraints to modify KCC to apply to this test.
- CGWL_10y_IPCC: we cannot look into the future and read what future editions of the IPCC will assess. CGWL10y_forec and CGWL10y_for_halfU. It was well beyond our available computational ability to run an ensemble of decadal forecasts initialized to each of the 65 years * 270 future test ensemble members.

2.3.2 Further particulars on the MPI-ESM1.2 LR SMILE ensemble

The MPI_ESM1.2 model version is described in Mauritsen et al. (2019) and references therein. The SMILE ensemble of MPI_ESM1.2 LR is run in the low resolution (LR) configuration (1.8° atmosphere, 1.5° ocean) and consists of 50 ensemble members for both historical period and each future core SSP scenario. Future scenario members are initialised from the end of the historical forcing runs at the end of 2014. Each future scenario is the standard SSP scenario configuration as deployed in the DECK scenario runs using this model. Further details on the configuration are given in Olonscheck et al. (2023). Figure S3 illustrates how the different ensembles in the MPI-ESM1.2 LR simulations are exhibiting divergent behaviour at the timing of exceeding 1.5°C of warming. Note that crossing times for 1.5°C are systematically somewhat later than in the IPCC AR6 assessment which assessed 1.5°C crossing to occur by the early 2030s for all scenarios, but for the purposes of method performance assessment this is not material per se for the question posed in the present paper. What matters for the present paper is the divergent trajectories around the timing of exceedance.

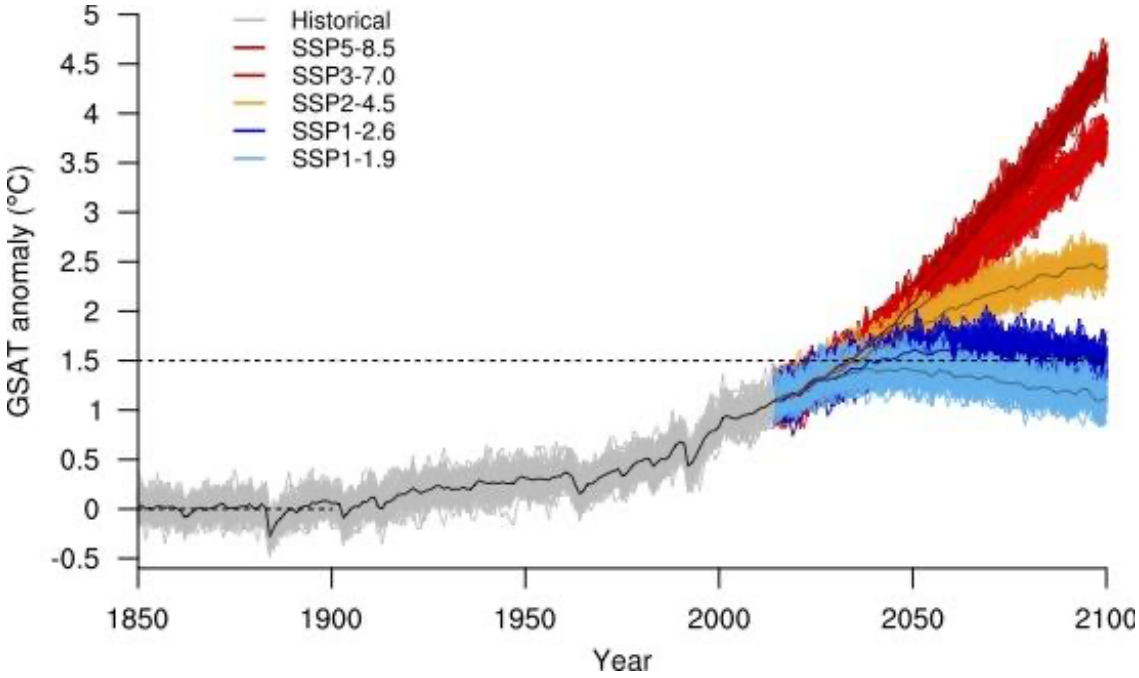

Figure S3. The MPI-ESM1.2-LR 50-member SMILE ensemble results for historical and the standard CMIP6 five future SSP scenarios run to 2100. SSP1-1.9 in the ensemble mean never reaches 1.5°C on a sustained basis. For SSP1-2.6 the ensemble mean exceeds and then returns towards 1.5°C by the end of the 21st Century. All remaining scenarios exceed 1.5°C by a substantial margin and are still warming by the end of the century. The switch to scenarios occurs in 2015 with the cessation of the historical forcings from CMIP.

2.3.3 Complete results for the MPI-ESM1.2 LR and NorESM1 Volc Tests

https://github.com/jnickla1/Thorne_15/blob/main/averaged_runsNorESM_copy.xlsx

https://github.com/jnickla1/Thorne_15/blob/main/averaged_runsMPIESM_copy.xlsx

We direct the readers to these two files, which also contain additional metrics as columns beyond Tables 3 and 4. These are detailed below. Note that additional columns "D-I" without headers in the average_runs245 tab in averaged_runsMPIESM_copy.xlsx were used to create and format Table 3 - they copy the 75RMS columns from other sheets. Highlighting within averaged_runsNorESM_copy.xlsx serves this same purpose. All metrics besides RMSEs were combined across all ensemble members with a simple average, whereas RMSEs were combined as mean squared errors before reapplying the square root.

100bias: bias evaluated relative to the 20-year running mean over both 2000-2090

bias50: bias evaluated over a shorter period 2050-2090.

bias: bias evaluated over a longer period 1850-2090.

#q<0.1: average # of q-values from 1850-2090 that are smaller than 0.1 (detectable difference from 20-year running mean)

#q<0.5: average # of q-values from 1850-2090 that are smaller than 0.5

100log-l: sum of log-likelihoods relative to 20-year running mean from 2000-2090

Edyrs15: most extreme difference in 1.5°C crossing time (in years) relative to the 20-year running mean (if there are multiple crossing times).

Fdyrs15: difference in the first 1.5°C crossing time relative to the 20-year running mean

Mdyrs15: median difference in the first 1.5°C crossing time relative to the 20-year running mean (computed from the equilibrium point of the PDF curves)

""dyrs20: same as above but for 2.0°C crossing times, defaults to -1 if the 20-year running mean never crosses.

EdyrsA: same as above but averaging all crossing times from 1.1°C up to the maximum reached by the 20-year running mean

RMSyrsA: RMSE evaluated over a longer period 1850-2090 relative to the 20-year running mean

100RMS: RMSE evaluated over 2000-2090 relative to the 20-year running mean

75RMS: RMSE evaluated over 2025-2090 relative to the 20-year running mean (used in main text tables)

l15: log-likelihood of 1.5°C in the year the 20-year running mean crosses 1.5°C

l20: log-likelihood of 2.0°C in the year the 20-year running mean crosses 2.0°C

log-likeli: average log-likelihood of the 20-year running mean over entire 1850-2090 period

ncEdyrs: number of times that a 0.1°C threshold is crossed from 1.1°C to the end of the record by that metric (up to the max temperature achieved by the 20-year running mean). So if the 20-year running mean reaches 1.65°C and the method is linear, it will have exactly 6 crossings.

nceEdyrs: Same as the above, but counting up to the maximum temperature reached by the ensemble average 20-year running mean (same evaluation for all ensemble members).

q_min: the minimum q-value, corresponds to the certainty we can detect one year differs from the 20-year mean

q_small5: : the 5th smallest q-value, corresponds to the certainty we can detect that a set of 5 years differ from the 20-year mean

smooth_r: the ratio of the unsmoothness of a particular method (entire 1850-20100 window) compared to the 20-year running mean unsmoothness (1860-2090). We define unsmoothness as the mean absolute 2nd differences (2nd derivative) in the central estimate. All straight lines have 0 unsmoothness.

tlog-l: the sum of the log-likelihoods of the 20-year running mean from 1860-2090. Methods that don't report anything for certain years generally are penalized by this metric compared to the average log-likelihood, as most methods have positive log-likelihood in most years.

e"": any of the above metrics but evaluated relative to the ensemble average 20-year running mean (rather than a within-member comparison).

2.3.4 Further sensitivity analysis results for the combination of different methods to estimate current long-term warming

Table S4 provides a more comprehensive overview of various sensitivity tests for the combination of various candidate techniques to estimate the present longterm warming level as described in Section 4.6.3 in the main text. Results are found to be relatively insensitive to a range of reasonable choices.

	First crossing instant of 1.5°C			Kullback-Liebler Divergence		# crosses	RMSE	
Ensemble	within 1 yr	within 2 yrs	within 5yrs	2000-2090	2025-2090	of 1.5°C	2000-2090	2025-2090
	Compressed Mixture (CSCM), 7 methods							
ESM1-2-LR_SSP370_constVolc	54.6%	85.0%	100.0%	0.352	0.271	1.080	0.0368	0.0335
ESM1-2-LR_SSP245_constVolc	55.7%	85.1%	99.2%	0.488	0.481	1.040	0.0322	0.0301
ESM1-2-LR_SSP126_constVolc	20.8%	39.3%	75.5%	0.568	0.602	1.460	0.0356	0.0348
NorESM_RCP45_VolcConst	53.3%	87.1%	100.0%	0.591	0.623	1.017	0.0300	0.0300
NorESM_RCP45_Volc	29.6%	54.3%	87.9%	0.711	0.771	1.100	0.0495	0.0510
	Compressed Mixture (CSCM), 18 methods							
ESM1-2-LR_SSP370_constVolc	55.7%	85.1%	100.0%	0.345	0.269	1.060	0.0365	0.0332
ESM1-2-LR_SSP245_constVolc	55.3%	83.5%	99.1%	0.474	0.465	1.080	0.0322	0.0300
ESM1-2-LR_SSP126_constVolc	20.6%	39.1%	76.0%	0.555	0.589	1.320	0.0354	0.0345

NorESM_RCP45_VolcConst	53.5%	86.7%	100.0%	0.568	0.475	1.017	0.0297	0.0295
NorESM_RCP45_Volc	30.1%	55.2%	88.5%	0.721	0.710	1.100	0.0492	0.0508
	Compressed Mixture (CSCM), 37 methods							
ESM1-2-LR_SSP370_constVolc	54.4%	84.6%	100.0%	0.356	0.272	1.060	0.0370	0.0335
ESM1-2-LR_SSP245_constVolc	53.8%	83.0%	99.3%	0.473	0.466	1.060	0.0319	0.0299
ESM1-2-LR_SSP126_constVolc	21.8%	41.5%	79.4%	0.554	0.589	1.320	0.0348	0.0341
NorESM_RCP45_VolcConst	54.5%	86.9%	100.0%	0.572	0.598	1.017	0.0294	0.0293
NorESM_RCP45_Volc	32.6%	59.1%	90.9%	0.736	0.814	1.083	0.0490	0.0507
	Inverse Variance (PIVW), 7 methods							
ESM1-2-LR_SSP126_constVolc	21.1%	39.9%	74.6%	0.529	0.578	1.420	0.0381	0.0383
ESM1-2-LR_SSP245_constVolc	51.3%	80.8%	98.6%	0.404	0.409	1.040	0.0357	0.0350
ESM1-2-LR_SSP370_constVolc	56.3%	85.9%	100.0%	0.281	0.215	1.120	0.0351	0.0319
NorESM_RCP45_Volc	30.8%	56.1%	88.3%	0.688	0.736	1.117	0.0527	0.0545
NorESM_RCP45_VolcConst	54.5%	87.5%	100.0%	0.437	0.468	1.000	0.0306	0.0310
	Inverse Variance (PIVW), 18 methods							

ESM1-2-LR_SSP126_constVolc	21.9%	41.4%	77.9%	0.487	0.514	1.320	0.0351	0.0336
ESM1-2-LR_SSP245_constVolc	50.2%	80.1%	98.2%	0.378	0.361	1.040	0.0338	0.0311
ESM1-2-LR_SSP370_constVolc	53.5%	84.8%	100.0%	0.301	0.206	1.040	0.0365	0.0309
NorESM_RCP45_Volc	31.1%	56.6%	88.3%	0.673	0.712	1.100	0.0515	0.0528
NorESM_RCP45_VolcConst	54.8%	87.6%	100.0%	0.424	0.446	1.000	0.0298	0.0295
	Inverse Variance (PIVW), 37 methods							
ESM1-2-LR_SSP126_constVolc	23.3%	43.8%	79.7%	0.491	0.535	1.420	0.0355	0.0354
ESM1-2-LR_SSP245_constVolc	53.1%	82.5%	99.0%	0.358	0.353	1.040	0.0322	0.0306
ESM1-2-LR_SSP370_constVolc	57.1%	86.9%	100.0%	0.294	0.238	1.060	0.0359	0.0335
NorESM_RCP45_Volc	30.7%	56.1%	88.5%	0.675	0.709	1.150	0.0515	0.0525
NorESM_RCP45_VolcConst	56.1%	88.2%	100.0%	0.419	0.437	1.017	0.0295	0.0288

Table S4: As in Table 5, but with more options for the number of methods included.

2.4 Description of generation of additional forcing and ocean temperature data (Sections 4.6 and 5)

For external forcing estimates (relevant for the FaIR and EBM-KF methods), we used a random 100 member subset (using simple rather than balanced k-means sampling) of the 841 ensemble members of net top-of-atmosphere forcing estimates from Smith *et al.* (2024). FaIR internally does a thorough job of assimilating the uncertainty in this external forcing ensemble (especially anthropogenic forcings), so we leave the ensemble as it is calibrated. While future expansion / modification of this code is possible, the EBM-KF-ta4 method currently designates anthropogenic and natural forcings as known inputs to an uncertain model, whereas the net top-of atmosphere forcing is an uncertain observation.

For ocean heat content estimates (relevant only for the EBM-KF-ta4 method), we generated 100 ensemble member time series (Figure S4) to represent its structural uncertainty using 5 instrumental ocean heat content datasets (IAPv4, Cheng *et al.*, 2024; CSIRO, Domingues *et al.*, 2008; EN4, Good *et al.*, 2013; JMA, Ishii *et al.*, 2017; NCEI, Levitus *et al.*, 2012), 1 community consensus estimate constructed from these 5 instrumental datasets (GCOS, von Shuckmann *et al.*, 2024), and 7 reanalysis datasets (CORA, Szekely *et al.*, 2019; MoHEACAN, Marti *et al.*, 2023; Minière *et al.*, 2023; the four GREP reanalyses, Cocetta *et al.*, 2024). Because these datasets do not have perfect coverage, we infilled data gaps by considering three different depth layers (0-700m, 700-2000m, 2000m-6000m) and different regions (e.g., latitudinal bands). For each year, ensemble member, depth range, and region, a selection is made among the available datasets, ensuring that each available dataset is sampled a proportionate number of times (details are provided in the zenodo dataset description); although, we considered the four variants of EN4 as a single dataset and the four GREP reanalyses as a single dataset for this sampling. We used Zanna *et al.* 2024 estimates when no other ocean heat content estimates were available, but otherwise avoided using these estimates since these estimates were inferred from sea surface temperature observations, so these estimates have modeled unconstrained transport biases and their inclusion could result in double counting of SST information. To extend Zanna *et al.* estimates prior to 1870, we assumed deterministic conditions were the same as in 1870, but with added AR1 noise. For each ensemble member, an associated non-structural uncertainty time series was estimated and treated as temporally uncorrelated between years.

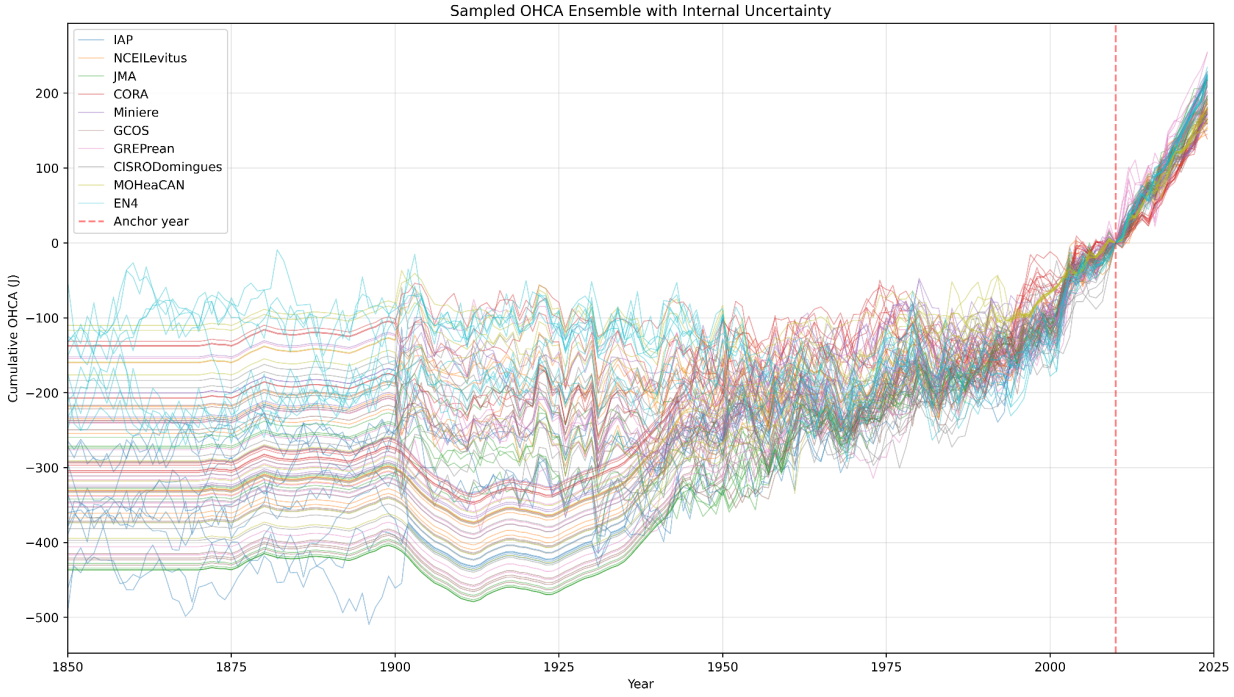

Figure S4. Annual OHCA (all-latitudes, all-depths) as a 100-member ensemble, constructed by a similar family tree method to the preceding figure for global mean temperature anomalies. In this case, the family tree draws available data region-by-region from the available reconstruction, reanalysis, and observational datasets (Supplemental Table S5). Additionally, AR1 process noise is

added to provide consistency of each ensemble member as a draw from the internal dataset uncertainty when the ensemble spread would underestimate the uncertainty. The Zanna et al. 2024 reconstruction is not used during the ARGO era, but from 1870-1900 it is the only product available. Before 1870, no net change in OHCA is assumed, but AR1 noise is used to broaden the ensemble spread consistent with the first uncertainty value from Zanna et al.

Next, for each of the 7 realized and 2 attributable methods of Section 4, we estimate realized and anthropogenic warming 100 times, once with each of the ensemble member time series of temperature anomalies, and for EMB-KF-ta4 also 100 sampled members of radiative forcing, ocean heat content, and ocean heat content non-structural uncertainty. When applying these Section 4 methods, we assume no time-varying, measurement-related uncertainty for each ensemble member time series of temperature anomalies and radiative forcing, which avoids a double counting of observational uncertainty. However, the bulk of the total uncertainty reported by each of these methods is intrinsic to each method and unrelated to the observational uncertainty. These intrinsic uncertainties generally arise from: the mechanics of nearest-neighbor sampling within a finite ensemble (FaIR, CGWL_10y_sfUKCP), solving for coefficients within an over-specified linear system (lowess, GWI), or core dynamical parameters of a Bayesian system (EBMKF).

Name	Citation	0-700m START	0-700m END	0-2000m START	0-2000m END	0-6000m START	0-6000m END	Regional Coverage
IAP	Cheng et al., 2024			1940	2024	1992	2024	global
NCEI/Levitus	Levitus et al., 2012	1955	2024	2005	2024			global
CORA	Szekely et al., 2019					1960	2024	global (gridded)
EN4.2.2.c14, c13, g10, l09	Levitus et al., 2009; Gouretski & Reseghetti, 2010;Cowley et al., 2013; Cheng et al., 2014 bias correction for c14					1900	2025	global (gridded)
Minière	Minière et al., 2023					1960	2023	global
GCOS	von Schuckmann et al., 2023					1960	2020	60N to 60S

GREP reanalyses	Cocetta et al., 2024			2005	2019			60N to 60S
JMA	Ishii et al., 2017			1955	2024			global
CSIRO/Domingues	Domingues et al., 2008	1950	2023					60N to 60S
MOHeaCAN	Marti et al., 2022					1993	2021	global

Table S5: Details of the different ocean heat content products used to form the ensemble estimates. Note that none of the datasets space the whole historical period. Von Schuckman et al. (2024) and Minière et al. (2023) were valuable resources in identifying these data.

These data products were combined so as each depth range and region was selected from one available dataset to provide complete data coverage. Then additional stochastic AR1 noise was added to the records where the spread from different combinations of products did not exceed the reported uncertainty. In this way the ensemble has at least as much aleatoric uncertainty as each reported record. The sampling from different datasets provides the estimate of the epistemic uncertainty to do with different modeling, reconstruction, and statistical methods–as well as inclusion of different observations–which is usually greater than the aleatoric uncertainty.

Systemic error for the EN4 dataset was established by fitting a function from the average pointwise uncertainty and the mean observation weight to the IAP dataset's uncertainty, which was also adjusted on per region. CORA provided measures of uncertainty that we did not find as useful, so the systemic uncertainty for each region and year was simply copied from EN4, the other gridded dataset. Some regions were set to a systemic uncertainty of 0, for instance the 600-2000m layer if the shallower 0-600m layer had more reported uncertainty than the whole 0-2000m chunk.

Once these records were downloaded and pre-processed, we constructed a 100-member ensemble of global ocean heat content anomaly (OHCA) timeseries spanning 1850-2024 by systematically combining observational products and their spatial/temporal subsets.

Starting from 2010 (the period of maximum observational alignment), we processed each year through three sequential steps: (1) latitude infilling, where records covering only 60°S-60°N were extended to global coverage by adding polar regions from complementary datasets (alternating between CORA and EN4 bias-correction variants); (2) depth infilling, where records covering 0-700m or 0-2000m were extended to full depth (0-6000m) using available deeper observations or the Zanna reconstruction; and (3) temporal infilling, where missing years were populated by cycling through all complete records available for that year. This process swept forward to 2024, then backward to 1850, with the Zanna reconstruction added to

the available pool during the backward sweep. Each ensemble member's provenance was tracked via notation strings recording all component datasets and their spatial domains.

We augmented the ensemble by sampling within-record uncertainties using an autoregressive AR(1) process. For each cell, we calculated AR(1) coefficients from detrended 41-year windows, then generated temporally-correlated standard normal deviates that evolved according to these coefficients while maintaining unit variance. These were scaled by reported standard errors (using skewed-normal transformations for asymmetric uncertainties) and added to cumulative OHCA timeseries, yielding 100 physically plausible ocean heat content realizations spanning observational and structural uncertainties.

3 Supplementary information for Section 5

Bayesian Interpretation

Our methodology has a Bayesian interpretation, whereby the assignment of probabilities to datasets, ocean heat content datasets, and section 4 methods corresponds to the assignment of hyperpriors to each dataset or section 4 method. Under this interpretation, the probability that exactly one of the GMST/GSAT datasets, exactly one of the ocean heat content datasets, and exactly one of the section 4 methods have the correct statistical models is one, which, while implausible, provides a reasonable framework to account for structural uncertainty. We consider the uncertainty distribution of each dataset as the posterior distribution given observations and conditional on that dataset having the correct statistical model. Some datasets have explicitly Bayesian frameworks (e.g., GETQUOCS), while others have explicitly frequentist frameworks (e.g., HadCRU_MLE). For frequentist datasets, their frequentist likelihood distributions can be interpreted as approximate Bayesian posterior distributions as their likelihood distributions are approximately multivariate normal and by invoking objective Jeffrey's priors. The application of the section 4 methods can be interpreted as a Bayesian update to account for the information of the section 4 methods and other data sources such as radiative forcing data. Therefore, the results of our study can be interpreted as posterior distributions, and our uncertainty intervals can be considered credible intervals.

Limitations of the dataset merging method

Annual global time series (for each ensemble member or for a best estimate if no ensemble was available) were used for each dataset, either as provided by the dataset provider or calculated by us from a simple mean of their monthly series or an annual mean of area-weighted averages of available grid cells if only monthly grids were available.

Ideally, each dataset would have a native ensemble that included uncertainties associated with all known sources of error as they pertained to that dataset. Unfortunately, this is not the case. By using perturbations calculated from NOAAGlobalTempv5.0, ERA5, and HadCRUT5 we may have mis-estimated the uncertainty. It's not clear whether it would

be an underestimate or an overestimate. However, using only the best estimates from non-ensemble datasets would have led to a clear underestimate of the uncertainty which is obviously undesirable (see sensitivity tests).

By splicing the tail and head time series together using the midpoint of the reference period, our approach effectively treats uncertainties prior to 1996 as uncorrelated with uncertainties after 1995. As a result, our approach might underestimate uncertainty in the long-term change in global temperature since 1850-1900. The linear correlation coefficient in temperature change from 1850-1900 to 2023 is negative (between -0.10 and -0.20) according to the ensemble members of HadCRUT5 Analysis (and related datasets), DCENT_MLE, and GloSAT.

By using a single hierarchical tree, we perhaps give a greater weight to certain commonalities than others. In practice, a more optimal weighting could be found based on the expected covariances between datasets. The covariances would reflect the different weights given to the input datasets and any commonalities between them. This would still require subjective choices to be made and the choices would be more numerous and more difficult. Furthermore, the covariances would likely vary in time which means that no single choice of dataset weighting would be correct for all time steps. Therefore, to keep things practical, the simpler family-tree method was retained.

Dataset merging sensitivity tests

A number of sensitivity tests were performed to see to what degree a range of reasonable methodological choices might affect the resulting long-term warming estimates, including their uncertainties. In most respects they did not matter greatly with minor differences between reasonable choices. The mean of the ensemble generally changed very little unless obviously poor choices were made (such as deliberately choosing an unbalanced ensemble). The spread of the ensemble varied more.

The choice of family tree was tested by using a range of different trees using different subsets of data as follows (Figure S5):

1. Grouping by SST dataset (a)
2. Grouping by LSAT dataset (b)
3. Grouping by interpolation method (c)
4. Equal weighting for all datasets (d)
5. Tree-of-trees method combining the trees 1-3 (not shown)
6. Random tree, with each ensemble member using a randomly generated tree. (not shown)
7. Deliberately unbalanced tree (e)
8. As for 1 but with reanalyses grouped with the most similar SST datasets (f)

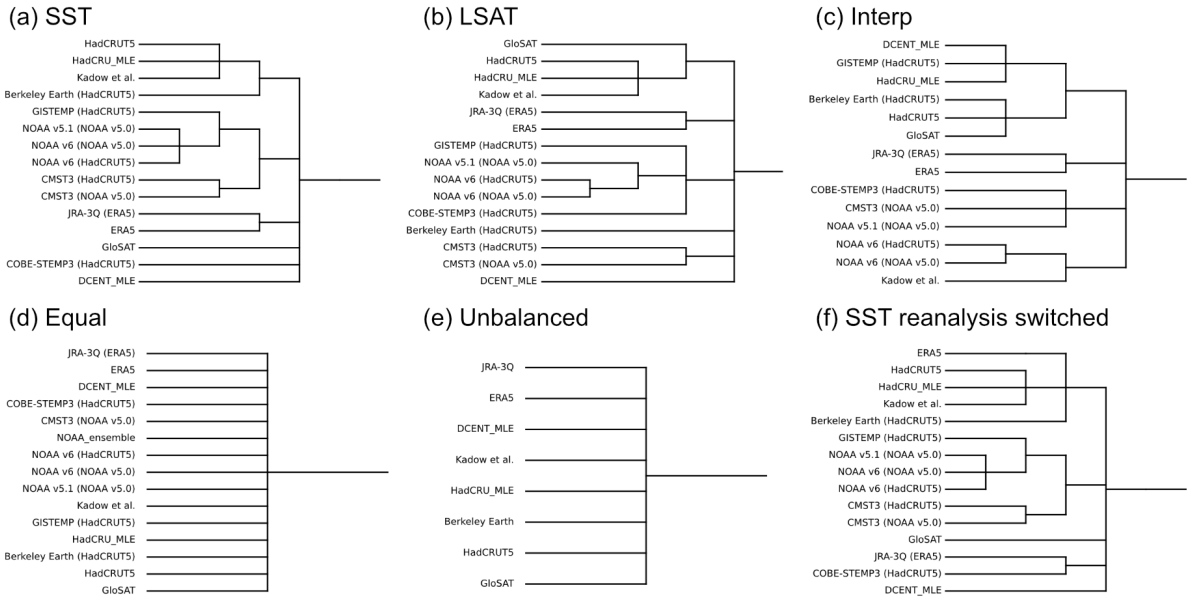

Figure S5. Alternative family tree approaches used to create the large multidataset ensemble. These distinct approaches yield different implicit weightings in ensemble construction to individual underlying products.

The LSAT grouping (2) generally had the largest ensemble spread (Figure S6), followed by the tree of trees (5) and SST(1) groupings. Equal weighting (4) and interpolation (3) trees gave similar spreads followed by random weighting (6), and finally the unbalanced tree (7) which was dead last (by design).

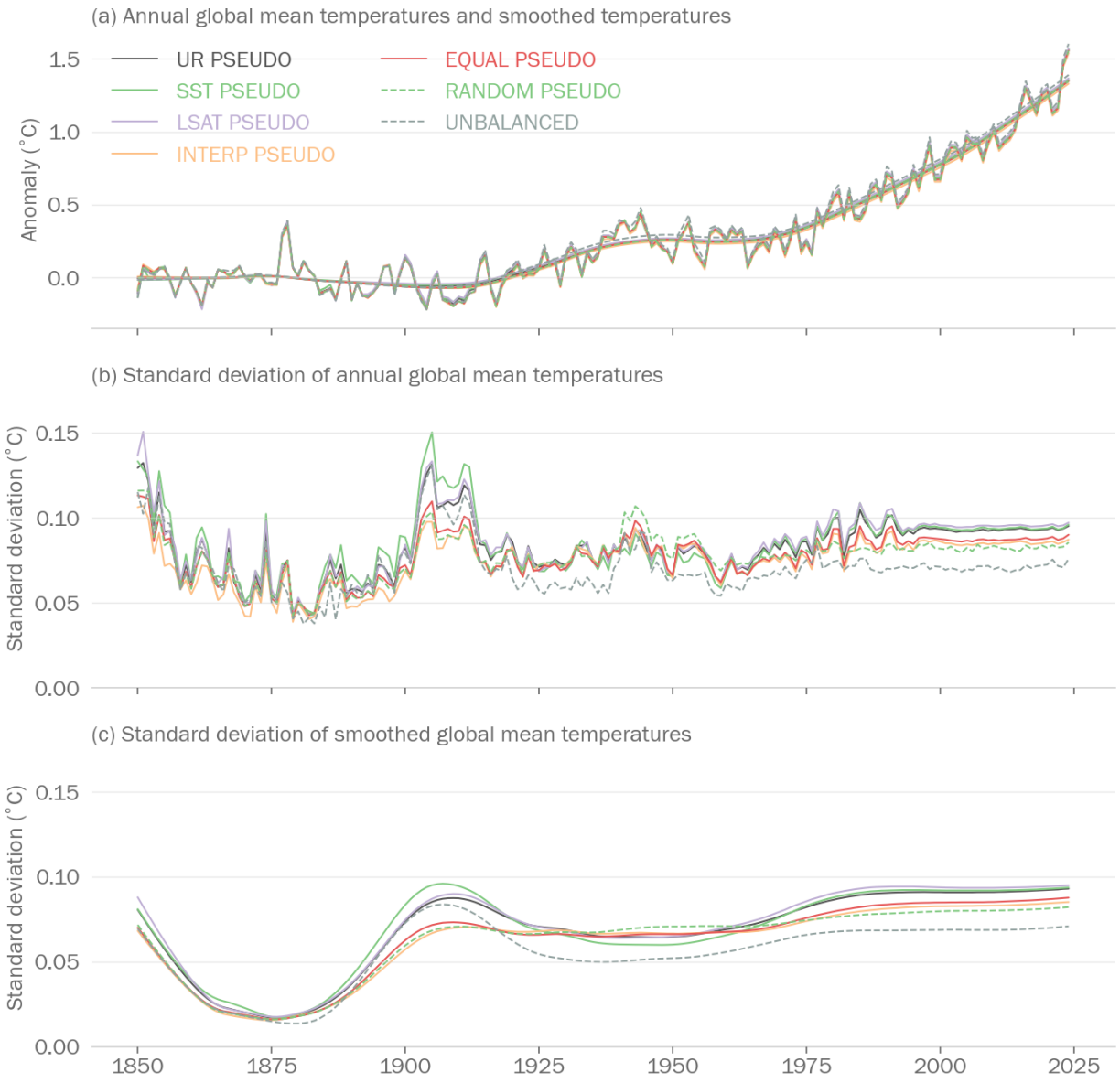

Figure S6: Comparison of the effect of different dataset family trees on (a) the global temperature ensemble mean; (b) the ensemble spread at annual timescales; and (c) the ensemble spread at multi-decadal timescales. The ensemble spread is represented by its standard deviation from its 1850-1900 average.

In addition, some of these were run in three modes:

1. Using datasets without modification (mix of ensemble and non-ensemble datasets).

2. Using native ensemble datasets only (hence a reduced set of datasets)
3. Using native ensemble datasets and pseudo ensembles generated for every dataset that did not have one.

Overall, methods which used ensembles (2, 3) typically had larger ensemble spreads (Figure S7) than those that did not (1). Methods using pseudo ensembles (3) had the largest spread after around 1930 and only native ensemble datasets (2) before. Ensemble-only datasets having a wider spread pre-1930 is likely due to large differences between GloSAT and DCENT_MLE in the early 20th Century though all choices have a larger ensemble spread at that time. Method (2) excludes all ERSST-based datasets so it warms more than the other combinations. All versions of NOAAGlobalTemp warm less than the balance of other datasets.

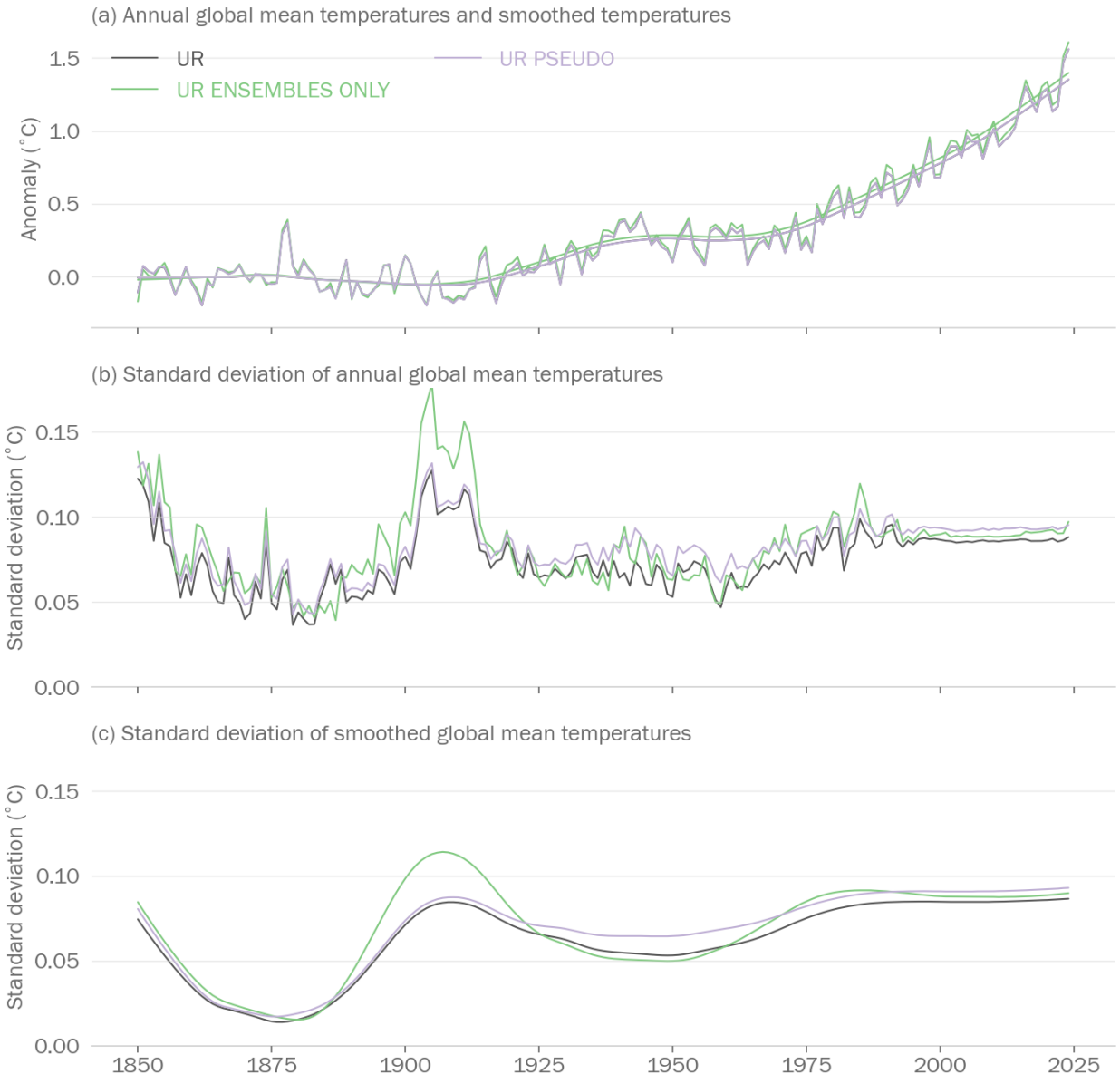

Figure S7: As Fig. S6, but comparison of non-ensemble, only native ensemble, and pseudo-ensemble datasets.

The choice to change datasets in the subperiod 1981-2010 was also explored (all overlaps were spliced at the midpoint).

1. 30-year fixed overlap, 1981-2010 (Basic)
2. 30-year overlap, any 30-year period common to the two datasets being spliced, and different each time (Variable overlap)
3. 10-year overlap, any 10-year period common to the two datasets being spliced, and different each time (Short overlap)

4. 2-year overlap, any 2-year period common to the two datasets being spliced, and different each time. (Shortest overlap)

435 The shortest overlap (4) generally had the largest ensemble spread (Figure S8) with longer moving overlaps (2, 3) having progressively smaller ensemble spreads. This is as one might expect as matching on a shorter period will be more affected by noise in single years and this will be averaged out on moving to longer periods. A fixed overlap (1) tends to have a comparable ensemble spread to that with the shortest overlap (4) except between roughly 1925 and 1980 where it is lower than the other methods (2-4).

440

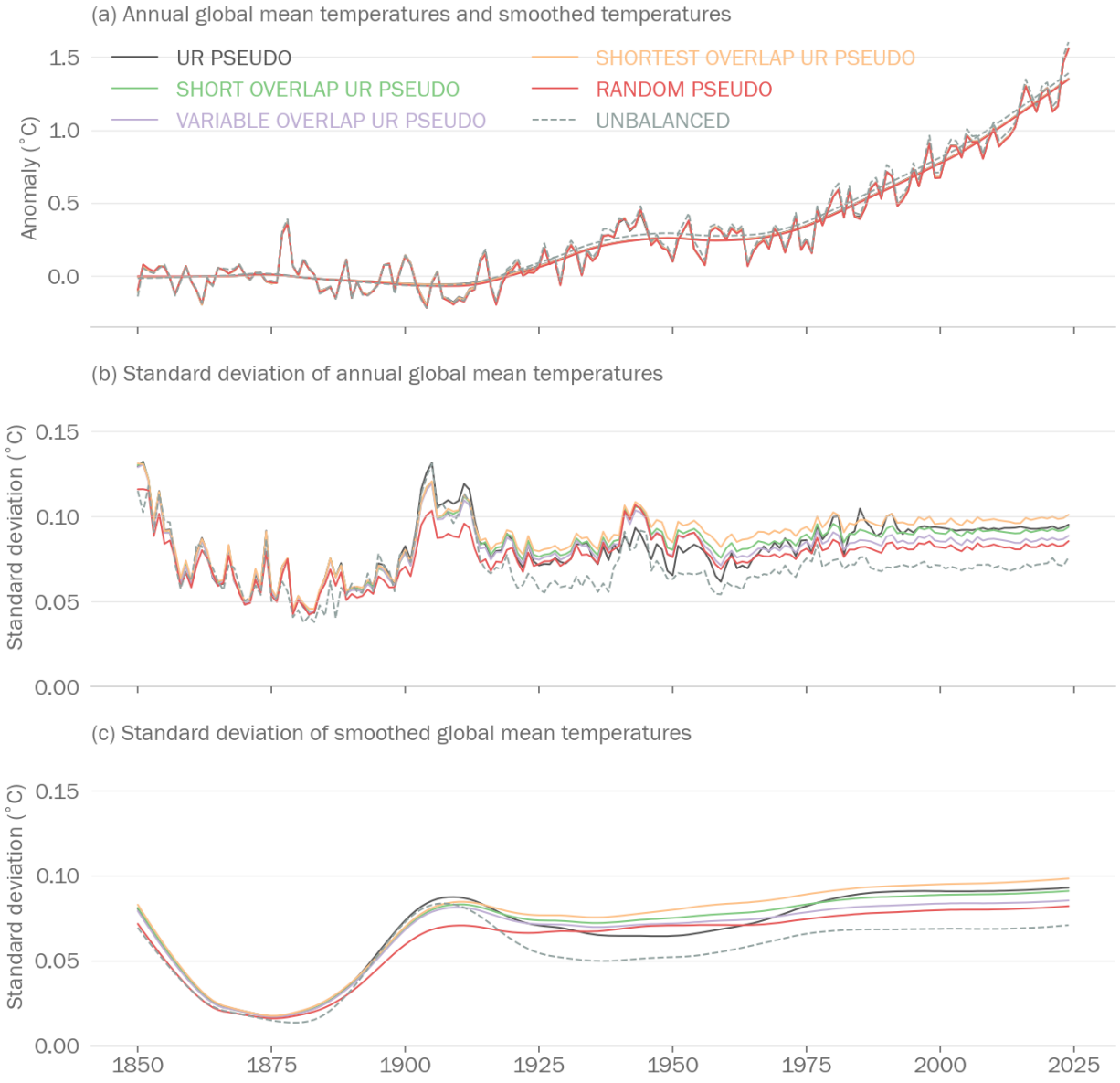

Figure S8: As Fig. S6, but comparison of different overlap splicing schemes.

We also compared 5 different methods of generating 100 ensemble members and evaluated their representativeness based on their Fréchet distances to a simple 10,000 member ensemble. In these calculations, we approximated distributions as multivariate normal using their sample means and covariance matrices. This metric considers differences in means, variances, and covariances when evaluating representativeness. We used Fréchet distances (Dowson & Landau, 1982) since many other common metrics used to compare multivariate normal distributions, such as Kullback-Leibler divergence (Kullback & Leibler, 1951), the Bhattacharyya distance (Schweppe, 1967), or the Hellinger distance (Eslinger & Woodward, 1990) would not be

applicable as relevant sample covariance matrices could have zero rank either due to a small number of ensemble members or due to our splicing of head and tail datasets.

The different methods of generating 100 ensemble members were:

1. simple random sample of 100 ensemble members
2. balanced k-means, in which balanced k-means was used to cluster the 10,000 member ensemble into 100 clusters each with 100 members and then one member from each cluster was selected at random
3. warming rate clustering, in which the ensemble members were ranked according to their long-term warming (difference between 1850-1900 and 2015-2024), split into 100 equal groups and then a random member of each group was chosen
4. balanced sampling, where 100 ensemble members were generated in such a way that the number of ensemble members generated from each dataset was representative of the underlying distribution (e.g., for the SST-based tree, an approximately equal number of ensemble members would be generated from each of the 3 HadCRUT5-based datasets)
5. rescaled ensemble, where we rescaled a simple random sample of 100 ensemble members to have the same sample mean and variances as the 10,000 member ensemble

Our results suggest that balanced k-means generated the most representative 100 member ensemble, although the reduction in the Fréchet distance was minor, perhaps due to the strong dependence on the HadCRUT5 ensemble limiting the effective dimensionality of the generated ensemble. Tests that used greater variation in donor ensemble datasets (not shown) found that balanced k-means, warming rate clustering, and balanced sampling greatly improved representativeness of the generated ensemble and even outperformed the simple random sampling of 1000 ensemble members. Techniques similar to our balanced k-means sampling have been used in diverse fields including chemical spectral data (Daszykowski *et al.*, 2002), cellular RNA data (Li *et al.*, 2022), and soil data (Robertson & Price, 2024).

The three methods all give similar results to the full ensemble (Figure S9) with the k-means method (2) perhaps coming closest of the three to the full ensemble though all means of reducing the ensemble are, as one might expect, noisier in the annual averages.

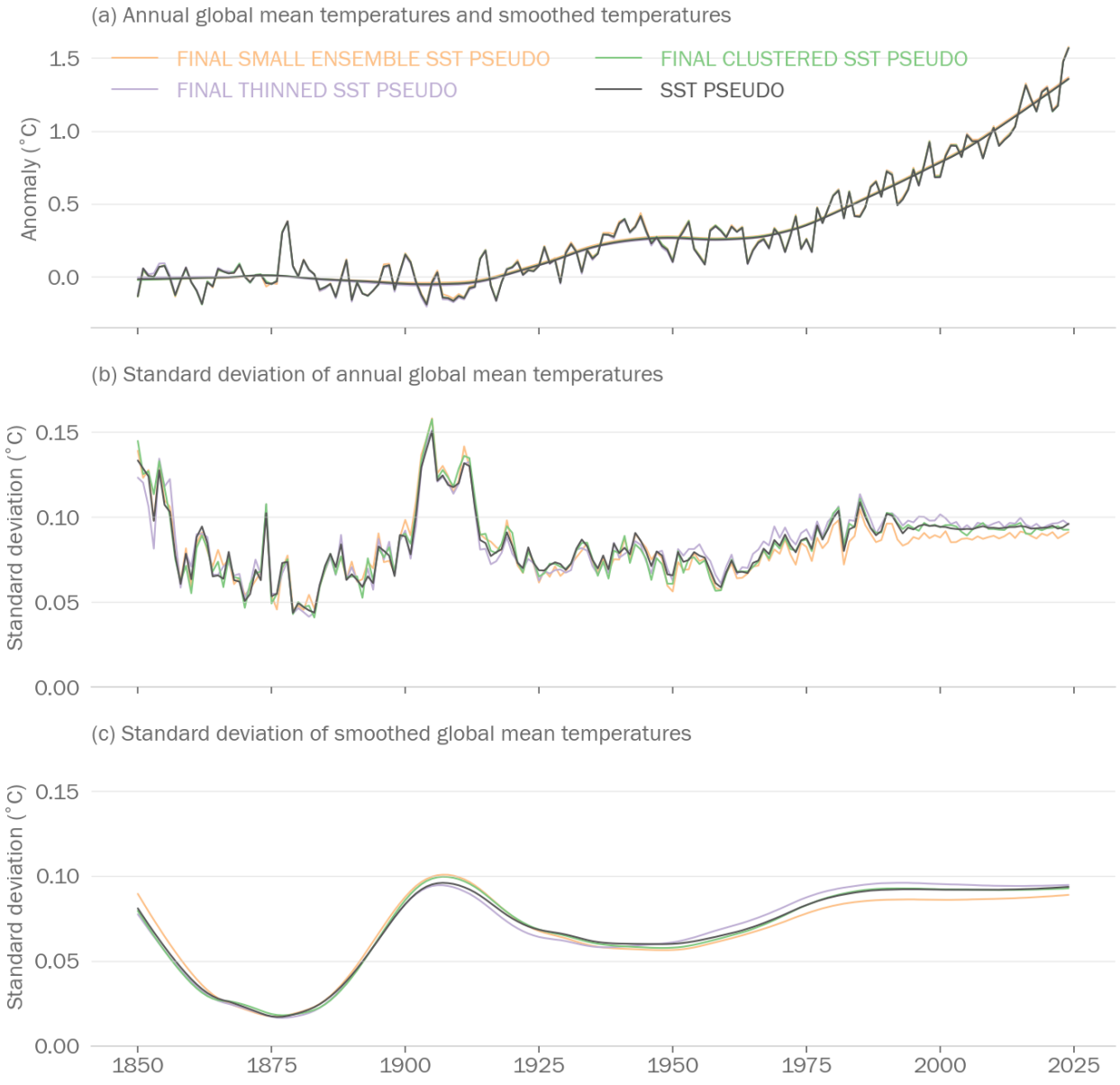

Figure S9: As Fig. S6, but comparison of different schemes for reducing the ensemble size.

Based on these sensitivity studies, it is suggested to generate:

- a large 10,000 member ensemble: which gives a stable estimate of the ensemble spread
- using a fixed overlap (1981-2010): preferred for operational reasons
- with each dataset converted to an ensemble if it was not already an ensemble: which makes use of the widest range of datasets and associated information
- The 10,000-member ensemble can be reduced to a 100-member ensemble using balanced k-means

References

Betts, R. A., et al.: Approaching 1.5 °C: how will we know we've reached this crucial warming mark?, Nature, 624, 33–35, https://doi.org/10.1038/d41586-023-03775-z, 2023.

Bevacqua, E., Schleussner, C.-F., and Zscheisler, J.: A year above 1.5 °C signals that Earth is most probably within the 20-year period that will reach the Paris Agreement limit, Nat. Clim. Change, https://doi.org/10.1038/s41558-025-02246-9, 2025.

Bône, C., Gastineau, G., Thiria, S., Gallinari, P., and Mejia, C.: Detection and attribution of climate change using a neural network, J. Adv. Model. Earth Syst., 15, e2022MS003475, https://doi.org/10.1029/2022MS003475, 2023.

Cannon, A. J.: Twelve months at 1.5 °C signals earlier than expected breach of Paris Agreement threshold, Nat. Clim. Change, https://doi.org/10.1038/s41558-025-02247-8, 2025.

Chen, X. and Tung, K. K.: Global-mean surface temperature variability: space–time perspective from rotated EOFs, Clim. Dyn., 51, 1719–1732, https://doi.org/10.1007/s00382-017-3979-0, 2018.

Cheng, L., J. Zhu, R. Cowley, T. Boyer, and S. Wijffels: Time, Probe Type, and Temperature Variable Bias Corrections to Historical Expendable Bathythermograph Observations. J. Atmos. Oceanic Technol., 31, 1793–1825, https://doi.org/10.1175/JTECH-D-13-00197.1., 2014

Clarke, D. C. and Richardson, M.: The benefits of continuous local regression for quantifying global warming, Earth Space Sci., 8, e2020EA001082, https://doi.org/10.1029/2020EA001082, 2021.

Cocetta, F., Zampieri, L., Selivanova, J. & Iovino D.: Assessing the representation of Arctic sea ice and the marginal ice zone in ocean–sea ice reanalyses. *The Cryosphere*, 18(10), 4687-4702. https://doi.org/10.5194/tc-18-4687-2024, 2024

Cowley, R., S. Wijffels, L. Cheng, T. Boyer, and S. Kizu: Biases in Expendable Bathythermograph Data: A New View Based on Historical Side-by-Side Comparisons. *J. Atmos. Oceanic Technol.*, 30, 1195–1225, https://doi.org/10.1175/JTECH-D-12-00127.1., 2013

Cummins, D. P., Stephenson, D. B., and Stott, P. A.: A new energy-balance approach to linear filtering for estimating effective radiative forcing from temperature time series, Adv. Stat. Clim. Meteorol. Oceanogr., 6, 91–102, https://doi.org/10.5194/ascmo-6-91-2020, 2020.

Daszykowski, M., Walczak, B. & Massart, D.L.: Representative subset selection. *Analytica Chimica Acta*, 468(1), 91-103. https://doi.org/10.1016/S0003-2670(02)00651-7, 2002.

Diffenbaugh, N. S. and Barnes, E. A.: Data-driven predictions of the time remaining until critical global warming thresholds are reached, Proc. Natl. Acad. Sci. USA, 120, e2207183120, https://doi.org/10.1073/pnas.2207183120, 2023.

Dowson, D.C. & Landau, B.V. :The Fréchet distance between multivariate normal distributions. *Journal of Multivariate Analysis*, 12(3), 450-455. https://doi.org/10.1016/0047-259X(82)90077-X, 1982.

Duan, Y., Kumar, S., and Kinter, J. L.: Evaluation of long-term temperature trend and variability in CMIP6 multimodel ensemble, Geophys. Res. Lett., 48, e2021GL093227, https://doi.org/10.1029/2021GL093227, 2021.

Eslinger, P.W. & Woodward, W.A.:Minimum hellinger distance estimation for normal models. *Journal of Statistical Computation and Simulation*, 39(1-2), 95-114. https://doi.org/10.1080/00949659108811342, 1990.

Foster, G. and Rahmstorf, S.: Global temperature evolution 1979–2010, Environ. Res. Lett., 6, 044022, https://doi.org/10.1088/1748-9326/6/4/044022, 2011.

Gillett, N. P., et al.: Constraining human contributions to observed warming since the pre-industrial period, Nat. Clim. Change, 11, 207–212, https://doi.org/10.1038/s41558-020-00965-9, 2021.

Gouretski, V. and Reseghetti, F.: On depth and temperature biases in bathythermograph data: Development of a new correction scheme based on analysis of a global ocean database. Deep Sea Research Part I: Oceanographic Research Papers, 57, 812-833, https://doi.org/10.1016/j.dsr.2010.03.011, 2010

Hartmann, D. L., Klein Tank, A. M. G., Rusticucci, M., Alexander, L. V., Brönnimann, S., Charabi, Y., Dentener, F. J., Dlugokencky, E. J., Easterling, D. R., Kaplan, A., Soden, B. J., Thorne, P. W., Wild, M., and Zhai, P.: Observations: Atmosphere and surface, in: Climate Change 2013: The Physical Science Basis, Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, edited by: Stocker, T. F., Qin, D., Plattner, G.-K., Tignor, M., Allen, S. K., Boschung, J., Nauels, A., Xia, Y., Bex, V., and Midgley, P. M., Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2013.

Hawkins, E. and Sutton, R.: The potential to narrow uncertainty in regional climate predictions, Bull. Am. Meteorol. Soc., 90, 1095–1108, https://doi.org/10.1175/2009BAMS2607.1, 2009.

Jarvis, A. and Forster, P. M.: Estimated human-induced warming from a linear temperature and atmospheric CO2 relationship, Nat. Geosci., 17, 1222–1224, https://doi.org/10.1038/s41561-024-01580-5, 2024.

Kullback, S. & Leibler, R.A. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1), 79-86. https://www.jstor.org/stable/2236703, 1951.

Lei, L., Lan, L.Y.-L., Huang, L., Ye, C., Andrade, J. & Wilson, P.C.: Selecting Representative Samples From Complex Biological Datasets Using K-Medoids Clustering. *Frontiers in Genetics*, 13, 954024. https://doi.org/10.3389/fgene.2022.954024, 2022.

Levitus, S., J. I. Antonov, T. P. Boyer, R. A. Locarnini, H. E. Garcia, and A. V. Mishonov: Global ocean heat content 1955–2008 in light of recently revealed instrumentation problems, *Geophys. Res. Lett.*, 36, L07608, doi:10.1029/2008GL037155, 2009.

Livezey, R. E., Vinnikov, K. Y., Timofeyeva, M. M., Tinker, R., and van den Dool, H. M.: Estimation and extrapolation of climate normals and climatic trends, J. Appl. Meteor. Climatol., 46, 1759–1776, https://doi.org/10.1175/2007JAMC1666.1, 2007.

Mann, M. E.: Smoothing of climate time series revisited, Geophys. Res. Lett., 35, L16708, https://doi.org/10.1029/2008GL034716, 2008.

Marti, F., Blazquez, A., Meyssignac, B., Ablain, M., Barnoud, A., Fraudeau, R., Jugier, R., Chenal, J., Larnicol, G., Pfeffer, J., Restano, M., and Benveniste, J.: Monitoring the ocean heat content change and the Earth energy imbalance from space altimetry and space gravimetry, Earth Syst. Sci. Data, 14, 229–249, https://doi.org/10.5194/essd-14-229-2022, 2022.

Mauritsen, T., Bader, J., Becker, T., Behrens, J., Bittner, M., Brokopf, R., et al.: Developments in the MPI-M Earth System Model version 1.2 (MPI-ESM1.2) and its response to increasing CO2, J. Adv. Model. Earth Syst., 11, 998–1038, https://doi.org/10.1029/2018MS001400, 2019.

Millar, R. J., Nicholls, Z. R., Friedlingstein, P., and Allen, M. R.: A modified impulse-response representation of the global near-surface air temperature and atmospheric concentration response to carbon dioxide emissions, Atmos. Chem. Phys., 17, 7213–7228, https://doi.org/10.5194/acp-17-7213-2017, 2017.

Minière, A., von Schuckmann, K., Sallée, J.B. and Vogt, L.: Robust acceleration of Earth system heating observed over the past six decades, Scientific Reports, 13(1), p.22975, https://doi.org/10.1038/s41598-023-49353-1, 2023.

Nicklas, J.M., Fox-Kemper, B. and Lawrence, C.: Efficient Estimation of Climate State and Its Uncertainty Using Kalman Filtering with Application to Policy Thresholds and Volcanism. J. Climate, 38(5):1235--1270. https://doi.org/10.1175/JCLI-D-23-0580.1, 2025.

Olonscheck, D., Suarez-Gutierrez, L., Milinski, S., Beobide-Arsuaga, G., Baehr, J., Fröb, F., et al.: The new Max Planck Institute Grand Ensemble with CMIP6 forcing and high-frequency model output, J. Adv. Model. Earth Syst., 15, e2023MS003790, https://doi.org/10.1029/2023MS003790, 2023.

Otto, F., Frame, D., Otto, A., et al.: Embracing uncertainty in climate change policy, Nat. Clim. Change, 5, 917–920, https://doi.org/10.1038/nclimate2716, 2015.

Qasmi, S. and Ribes, A.: Reducing uncertainty in local temperature projections, Sci. Adv., 8, eabo6872, https://doi.org/10.1126/sciadv.abo6872, 2022.

Robertson, B. & Price, C.: One point per cluster spatially balanced sampling. *Computational Statistics & Data Analysis*, 191, 107888. https://doi.org/10.1016/j.csda.2023.107888, 2024.

Samset, B. H., Zhou, C., Fuglestvedt, J. S., et al.: Steady global surface warming from 1973 to 2022 but increased warming rate after 1990, Commun. Earth Environ., 4, 400, https://doi.org/10.1038/s43247-023-01061-4, 2023.

Schweppe, F.C.: On the Bhattacharyya distance and the divergence between Gaussian processes. *Information and Control*, 11(4), 373-395. https://doi.org/10.1016/S0019-9958(67)90610-9, 1967.

Shumway, R. H.: Time series analysis and its applications, 4th edn., Springer Texts in Statistics, Springer International Publishing AG, 562 pp., https://doi.org/10.1007/978-3-319-52452-8, 2017.

Sippel, S., Meinshausen, N., Merrifield, A., Lehner, F., Pendergrass, A. G., Fischer, E., and Knutti, R.: Uncovering the forced climate response from a single ensemble member using statistical learning, J. Clim., 32 (17), 5677–5699, https://doi.org/10.1175/JCLI-D-18-0882.1, 2019.

Smith, C. J., Forster, P. M., Allen, M., Leach, N., Millar, R. J., Passerello, G. A., and Regayre, L. A.: FAIR v1.3: a simple emissions-based impulse response and carbon cycle model, Geosci. Model Dev., 11, 2273–2297, https://doi.org/10.5194/gmd-11-2273-2018, 2018.

Smith, C., Cummins, D. P., Fredriksen, H.-B., Nicholls, Z., Meinshausen, M., Allen, M., Jenkins, S., Leach, N., Mathison, C., and Partanen, A.-I.: fair-calibrate v1.4.1: calibration, constraining, and validation of the FaIR simple climate model for reliable future climate projections, Geosci. Model Dev., 17, 8569–8592, https://doi.org/10.5194/gmd-17-8569-2024, 2024.

Szekely, T., Gourrion, J., Pouliquen, S. & Reverdin, G.: The CORA 5.2 dataset for global in situ temperature and salinity measurements: data description and validation. *Ocean Science*, 15(6), 1601-1614. https://doi.org/10.5194/os-15-1601-2019, 2019

Trewin, B.: Assessing internal variability of global mean surface temperature from observational data and implications for reaching key threshold, J. Geophys. Res.-Atmos., 127, e2022JD036747, https://doi.org/10.1029/2022JD036747, 2022.

Visser, H., Dangendorf, S., van Vuuren, D. P., Bregman, B., and Petersen, A. C.: Signal detection in global mean temperatures after "Paris": an uncertainty and sensitivity analysis, Clim. Past, 14, 139–155, https://doi.org/10.5194/cp-14-139-2018, 2018.

von Schuckmann, K., Moreira, L., Cancet, M., Gues, F., Autret, E., Baker, J., Bricaud, C., Bourdalle-Badie, R., Castrillo, L., Cheng, L., Chevallier, F., Ciani, D., de Pasual-Collar, A., De Toma, V., Drevillion, M., Fanelli, C., Garric, G., Gehlen, M., Giesen, R., Hodges, K., Iovino, D., Jandt-Scheelke, S., Jansen, E., Juza, M., Karagali, I., Lavergne, T., Masina, S., McAdam, R., Minière, A., Morrison, H., Panteleit, T.R., Pisano, A., Pujol, M.-I., Stoffelen, A., Thual, S., Van Gennip, S., Veillard, P., Yang, C., and Zuo, H.: The state of the global ocean, State of the Planet, 4, 1-30, https://doi.org/10.5194/sp-4-osr8-1-2024, 2024.

Wills, R. C., Schneider, T., Wallace, J. M., Battisti, D. S., and Hartmann, D. L.: Disentangling global warming, multidecadal variability, and El Niño in Pacific temperatures, Geophys. Res. Lett., 45, 2487–2496, https://doi.org/10.1002/2017GL076327, 2018.

Wills, R. C. J., Battisti, D. S., Armour, K. C., Schneider, T., and Deser, C.: Pattern recognition methods to separate forced responses from internal variability in climate model ensembles and observations, J. Climate, 33, 8693–8719, https://doi.org/10.1175/JCLI-D-19-0855.1, 2020.

Wu, Z., Huang, N. E., Wallace, J. M., et al.: On the time-varying trend in global-mean surface temperature, Clim. Dyn., 37, 759–773, https://doi.org/10.1007/s00382-011-1128-8, 2011.

Yu, M. and Ruggieri, E.: Change point analysis of global temperature records, Int. J. Climatol., 39, 3679–3688, https://doi.org/10.1002/joc.6042, 2019.