PNAS
www.pnas.org

Supplementary Information for

Quantifying stochastic uncertainty in detection time of human-caused climate signals

B.D. Santer, J.C. Fyfe, S. Solomon, J.F. Painter, C. Bonfils, G. Pallotta, M.D. Zelinka

Corresponding Author: Benjamin Santer.
E-mail: santer1llnl.gov

This PDF file includes:

Figs. S1 to S7
References for SI reference citations

Statistical Analysis.

Method for correcting TMT data. All simulated and observed TMT results in the main text and SI are corrected for the contribution TMT receives from the cooling of the lower stratosphere. As in ref. (1), we used $\mathrm{TMT}_c = a_{24}\mathrm{TMT} + (1 - a_{24})\mathrm{TLS}$, where $a_{24} = 1.1$ and TMT_c denotes corrected TMT data. This method has been validated with both observed and model atmospheric temperature data (2–4).

Regridding, masking, and weighting. All satellite and model temperature datasets used in the fingerprint analysis are transformed from their original grid to a common $10^\circ \times 10^\circ$ latitude/longitude grid. After regridding, datasets are masked with the maximum common coverage in the observations. The latitudinal extent of regridded data is 80°N-80°S for TLS and TMT and 80°N-70°S for TLT. All model and observational temperature data used in the fingerprint analysis were appropriately area-weighted. Weighting involves multiplication by the square root of the cosine of the grid node's latitude (5).

Fingerprint calculation. Detection methods generally require an estimate of the true but unknown climate-change signal in response to an individual forcing or set of forcings (6). This is often referred to as the fingerprint $F(x)$. The fingerprint can be defined in different ways. A common strategy, which we employ here, is to define $F(x)$ as the first EOF of the ensemble-mean change in synthetic MSU temperature in either the CanESM2 or CESM1-CAM5 ALL simulations. Other possible fingerprint choices include the mean change or leading EOF from an equilibrium response experiment, a multi-model average trend pattern over some stipulated period of time, *etc.*

Let $S(i, x, t)$ represent annual-mean synthetic satellite temperature data at grid-point x and year t from the i^{th} realization of either the CanESM2 or CESM1-CAM5 ALL simulation, where:

$i = 1, \ldots N_r$ (the number of ALL realizations)
$x = 1, \ldots N_x$ (the total number of grid-points)
$t = 1, \ldots N_t$ (the time in years)

Here, N_r is 50 for CanESM2 and 40 for CESM1-CAM5. After transforming synthetic MSU temperature data from each model's native grid to the common $10^\circ \times 10^\circ$ latitude/longitude grid used for the fingerprint analysis, and after masking the regridded model data with observational coverage, $N_x = 576$ grid-points for TLS and corrected TMT, and 540 grid-points for TLT.

Model fingerprints are estimated over 1979 to 2018, the same time period used for determining observed temperature changes; N_t is 40 years.

The ensemble-mean atmospheric temperature change, $\overline{S}(x, t)$, was calculated by averaging over the total number of CanESM2 or CESM1-CAM5 ALL realizations. The overbar denotes this averaging step. Anomalies were then defined at each grid-point x and year t with respect to the local climatological annual mean. $F(x)$ is the leading EOF of the anomalies of $\overline{S}(x, t)$. Fingerprint patterns estimated from the CanESM2 ALL simulation are shown in Fig. S4.

In other studies, $F(x)$ is often rotated in a direction that maximizes the signal strength relative to the control run noise (6). While we have used optimized fingerprints in other work (7), no optimization of $F(x)$ was performed here.

Noise estimates. We seek to determine whether the pattern similarity between the time-varying observations and $F(x)$ shows a statistically significant increase over time. To address this question, we require estimates of internally generated variability in which we know *a priori* that there is no expression of the fingerprint (except by chance).

We obtain internal variability estimates from a multi-model ensemble of control runs, from the between-realization variability of two time periods of the CanESM2 ALL ensemble, and from the between-realization variability of the CanESM2 SV ensemble. We refer to these estimates subsequently as CMIP5, ALL1, ALL2, and SV.

In CMIP5, we rely on estimates of the internal variability of synthetic satellite temperature from 36 different CMIP5 pre-industrial control runs. We first define annual-mean temperature anomalies relative to climatological annual means over the full length of each control run. Because the length of the 36 control runs analyzed here varies by a factor of up to 4, models with longer control integrations could have a disproportionately large impact on our noise estimates. To guard against this possibility, our CMIP5 noise estimate relies on the last 200 years of each model's pre-industrial control run. Use of the last 200 years reduces the contribution of initial residual drift to noise estimates. Additionally, we assume that any drift behavior in the last 200 years can be well-approximated by a least-squares linear trend. The trend over the last 200 years is removed at each grid-point. The detrended annual-mean anomaly data are then concatenated.

In the ALL1 internal variability estimate, we calculate the ensemble-mean annual-mean temperature signal from CanESM2 over 1950 to 2100. This signal is subtracted from each of the 50 individual ALL realizations, and the residuals are then concatenated. ALL2 is defined similarly, but for 1979 to 2018 only. In SV, concatenation involves the residuals remaining after subtracting the ensemble-mean CanESM2 SV signal over 1950 to 2020 from each SV realization. The length of the CMIP5, ALL1, ALL2, and SV datasets is 7200, 7550, 2000, and 3550 years, respectively. Values of atmospheric temperature from the four noise estimates are regridded to the same $10^\circ \times 10^\circ$ target grid used for fingerprint estimation.

Projections onto the fingerprint. Observed annual-mean temperature data are first regridded to the same $10^\circ \times 10^\circ$ latitude/longitude grid used for the model ALL simulations and the four noise estimates, and are expressed as anomalies relative to climatological means over 1979 to 2018. The observed temperature data are then projected onto $F(x)$:

$$Z_o(t) = \sum_{x=1}^{N_x} O(x,t)\, F(x) \quad t = 1, \ldots, 40 \qquad [1]$$

where $O(x,t)$ denotes the observed annual-mean temperature data. This projection is equivalent to a spatially uncentered covariance between the patterns $O(x,t)$ and $F(x)$ at year t. The time series $Z_o(t)$ provides information on the fingerprint strength in the observations. If observed patterns of temperature change are becoming increasingly similar to $F(x)$, $Z_o(t)$ should increase over time.

In the analogous "model only" calculations, we search for a fingerprint estimated from either the CanESM2 or the CESM1 ALL ensemble in each individual ALL realization of that model. This involves projection of $S(i,x,t)$ onto $F(x)$:

$$Z_s(i,t) = \sum_{x=1}^{N_x} S(i,x,t)\, F(x) \quad i = 1, N_r\ ; \quad t = 1, \ldots, 40 \qquad [2]$$

where $N_r = 50$ for CanESM2 and $N_r = 40$ for CESM1.

Two approaches can be used to assess the significance of the changes in $Z_o(t)$ and $Z_s(t)$: direct comparison of test statistic values with some null distribution, or comparison of trends in $Z_o(t)$ and $Z_s(t)$ with a null distribution of trends. We use the trend approach here. To assess trend significance we use the four different noise data sets described above. These are denoted here by $C_1(x,t)$, $C_2(x,t)$, *etc.* The time series $N_1(t)$, for example, is the projection of $C_1(x,t)$ (the CMIP5 multi-model noise estimate) onto the fingerprint:

$$N_1(t) = \sum_{x=1}^{N_x} C_1(x,t)\, F(x) \quad t = 1, \ldots, 7200 \qquad [3]$$

In the following, we refer to $Z_o(t)$ and $Z_s(i,t)$ as "signal" time series, and we refer to $N_1(t), N_2(t)$, *etc.* as "noise" time series. It is these signal and noise time series that we use for calculating S/N ratios.

Estimating detection time. To calculate the timescale-dependent S/N ratios we use for estimating detection time t_d, we fit L-year least-squares linear trends to each signal time series. The start date of signal trends is 1979. The first signal trend is 10 years in length (1979 to 1988). L increases in increments of one year; the final 40-year signal trend is over 1979 to 2018. The signal trend is the numerator of the S/N ratio. To obtain the denominator of the ratio, non-overlapping L-year trends in pattern similarity are calculated from the noise time series. The denominator is the standard deviation of the distribution of these L-year noise trends.

Signal detection occurs at the trend length $L_{\rm d}$ for which S/N first exceeds the stipulated 3σ threshold and then remains continuously above the threshold for all values of $L > L_{\rm d}$. The detection time t_d is the final year of $L_{\rm d}$. We linearly interpolate between t_d and $t_d - 1$ in order to express the threshold exceedance time in fractional form.

Detection time results for a 5σ significance threshold are shown in SI Fig. S7. As expected, choice of a more stringent 5σ detection threshold delays fingerprint detection in both the "model only" results and in satellite data. Even with a 5σ threshold, however, the CanESM2 and CESM1 ALL fingerprints are still identifiable by 2018 in 87% of the comparisons with satellite TMT and TLT data.

Calculation of P_1. P_1 is the percentage of the total number of observational t_d values that lies within the stochastic uncertainty $t_{d\{r\}}$. For tropospheric temperature fingerprints, each of the two models has 60 different fingerprint detection times in satellite data: 36 for TMT (3 domains × 4 noise estimates × 3 satellite data sets) and 24 for TLT (3 domains × 4 noise estimates × 2 satellite data sets). For CanESM2, for example, there are 36 out of 60 cases in which the fingerprint detection time in satellite data lies within the stochastic uncertainty inferred from the CanESM2 ALL ensemble, so $P_1 = 60\%$. For TLS, each model has 36 fingerprint detection times in observations.

Fingerprint choice. Our S/N analysis relies on the CanESM2 and CESM1 ensemble-mean ALL fingerprints for TLS, TMT (corrected for lower stratospheric cooling), and TLT. Neither model was used to perform a large ensemble with anthropogenic forcing only. For CanESM2, however, the climate response to anthropogenic forcing (henceforth "ANT") can be estimated by subtracting the ensemble-mean SV signal (for each grid-point and year) from each individual ALL realization. As in the case of the CanESM2 ALL fingerprints, the ANT fingerprints were calculated over the 40-year period from 1979 to 2018.

For tropospheric temperature changes, the CanESM2 ALL and ANT fingerprint patterns are very similar, and primarily reflect the large warming signal arising from human-caused increases in well-mixed GHGs. For lower stratospheric temperature, there are small differences between the ANT and ALL fingerprint patterns. These differences arise because the large warming signals after the eruptions of El Chichón in 1982 and Pinatubo in 1991 occur in the first half of the satellite record, and therefore contribute to a negative trend in TLS over the full satellite era (Fig. 1A). Subtracting SV from ALL removes these volcanic signals and decreases overall lower stratospheric cooling, but does not markedly change the TLS fingerprint pattern. In practice, whether we use the ALL or ANT fingerprint pattern has relatively small impact on the S/N results shown here.

Differences between ToE and fingerprint results. Previous studies have shown that LEs are a valuable resource for estimating the local "Time of Emergence" (ToE) of an anthropogenic signal (8–12). ToE calculations differ from the detection time results presented here in two important ways. First, ToE is estimated locally, while our t_d results are based on forced and unforced spatial patterns of climate change. Use of pattern information is generally more efficient for partitioning signal and noise (6). Second, local ToE calculations consider how the amplitude of a forced climate signal evolves as the analysis period L increases, but estimate the size of natural climate variability over a single L-year timescale only (9, 13, 14). In our fingerprint method, the denominator of the S/N ratio used for determining t_d is estimated on the same timescale as the evolving signal, and we account for decreases in noise amplitude with increasing timescale (Figs. 4C,D).

Because of these differences, we do not expect the spatial average of local ToE results at thousands of grid-points to be directly comparable to our pattern-based t_d values.

Initialization of individual CanESM2 realizations. In each LE except GHG, the initial conditions in 1950 are taken from the five CanESM5 realizations submitted to the CMIP5 archive. The five realizations were branched into 50 realizations per experiment by introducing a random permutation to a seed used in the random number generator for the cloud physics scheme (15). This small change ensures that internal variability in the realizations diverges within months in the atmosphere and within years in the ocean. The forcing in each LE is identical, and the runs differ only in their realization of internal variability.

Analysis of CanESM2 ensembles with individual forcings. The CanESM2 LEs with individual forcings provide insights into the main drivers of atmospheric temperature change in the CanESM2 ALL simulation. Over 1979 to 2000, ozone depletion is the primary cause of lower stratospheric cooling. Smaller cooling contributions arise from increases in well-mixed GHGs and from volcanic forcing (SI Fig. S1). The volcanic effect is due to the selected 1979 to 2000 period and the asymmetric temporal distribution within this 22-year period of large stratospheric warming signals caused by El Chichón and Pinatubo (Fig. 1A). Over 2001 to 2018, partial recovery of stratospheric ozone (16, 17) leads to small, non-significant lower stratospheric warming (SI Fig. S2).

Forcing by increases in well-mixed GHGs is the main cause of significant tropospheric warming (SI Figs. S1, S2). This GHG signal is augmented in the early 21st century by small, non-significant warming arising from ozone recovery and from gradual "rebound" of upper ocean temperature after Pinatubo-induced cooling (SI Fig. S2) (18).

Credibility of model noise estimates. In the real world, we cannot directly observe 'pure' internal variability. Observed climate records are simultaneously influenced by both internal noise (operating on a wide range of different space and timescales) and by the signals arising from multiple anthropogenic and natural forcings. Because of this complex mix of signal and noise, the relatively short observational record lengths (particularly for satellite temperature data), and interactions between external forcing and internal variability (19, 20), there will always be some irreducible uncertainty in partitioning observations into internally generated and externally forced components. This uncertainty hampers assessment of how well current climate models capture the amplitudes, structures, and timescales of observed modes of internal variability (21).

In fingerprint detection work, we are most concerned about the possibility that individual models or a multi-model ensemble systematically underestimate observed variability, particularly on multidecadal timescales. Such an error would inflate S/N ratios and could lead to spurious detection of an anthropogenic fingerprint. Our previous work with multi-model ensembles yields the opposite result: on average, CMIP3 and CMIP5 models overestimate the amplitude of observed global-mean TMT variability on timescales of 10-15 years* (22, 23). This suggests that the detection times obtained here with the CMIP5 multi-model noise estimates are probably too conservative rather than too liberal.

The same applies to the LE-based estimates of tropospheric temperature variability obtained from CanESM2. In Santer et al. (2018), we found that the amplitude of decadal temperature variability for global-mean TMT (corrected for stratospheric cooling) is roughly a factor of 2-3 larger in CanESM2 than in satellite data (24). Further work is currently underway to evaluate the causes and statistical significance of differences between model and observed tropospheric temperature spectra.

Relevant regression-based studies. As noted in the main text, fingerprint detection times in satellite tropospheric temperature data are more consistent with the range of t_d values in the CESM1-CAM5 LE than with the corresponding range inferred from the CanESM2 LE. One interpretation of this finding is that CESM1-CAM5 has a more realistic estimate of the true tropospheric temperature signal caused by combined anthropogenic and natural external forcing. A second interpretation is that the closer agreement between t_d values in satellite tropospheric temperature data and in CESM1-CAM5 is primarily a consequence of error compensation. Under this second interpretation, ECS and negative anthropogenic aerosol forcing in CESM1-CAM5 are both larger than their true (but uncertain) real-world values.

Our analysis cannot discriminate between these two interpretations. At the time our research was performed, we did not have all of the single forcing simulations we would require in order to determine whether CESM1-CAM5 or CanESM2 provides a more credible estimate of the GHG and anthropogenic aerosol components of tropospheric temperature change. CESM1-CAM5 does not have "GHG only" and "AERO only" LEs. Both LEs are available from CanESM2, although the GHG signal must be estimated by subtraction (see Methods).

For CanESM2, therefore, it is possible to determine whether the "AERO only" and "GHG only" fingerprints are consistent in magnitude with observational climate data. Such analyses typically rely on calculating β, the regression coefficient between the spatio-temporal fingerprint pattern and observations (25). For example, a GHG fingerprint with a β value significantly greater

*Given that satellite temperature records are only 40 years long, 15-20 years is the longest timescale for which meaningful comparisons between simulated and observed TMT variability can be made.

than zero and consistent with 1 would indicate that the fingerprint had been detected in observations and was attributable to GHG forcing. A β value significantly greater than zero and significantly larger (smaller) than 1 would signify that the model-predicted fingerprint was detectable but weaker (stronger) than in observations.

The pattern-based regression study most relevant to our own investigation is by Lott et al. (2013) (26). Figure 9 in Lott et al. shows estimated β values for the model-predicted GHG, OAnt, and NAT components of tropospheric temperature change in 8 different CMIP5 models and in 4 different radiosonde datasets[†]. The OAnt fingerprint comprised all anthropogenic forcings except well-mixed GHGs[‡], while NAT included volcanic and solar forcing only. The CanESM2 results in Lott et al. showed that the GHG component of tropospheric temperature change was detectable in all four radiosonde datasets, but with an average β value that was smaller than one – i.e., the amplitude of the model GHG fingerprint had to be down-weighted in order to best match the observed spatio-temporal patterns of temperature change (26).

A second relevant paper by Swart et al. (2018) (15) used the CanESM2 ALL, GHG, NAT, OZONE, and AERO LEs to explore the detectability of the model-predicted fingerprints in observed hydrographic profiles of ocean temperature and salinity. For the GHG fingerprint, β values for both temperature and salinity were always greater than zero but significantly less than 1[§]. As in the Lott et al. (2013) analysis of tropospheric temperature, therefore (26), the CanESM2 GHG fingerprint in ocean temperature and salinity was robustly detectable in observations but had to be down-weighted to fit the observations. A similar result was found by Gillett et al. (2013) (27) for the CanESM2 GHG fingerprint in observed surface temperature records[¶].

These three regression-based studies do not yield consistent β values for the CanESM2 OAnt and AERO fingerprints. In Lott et al. (2013), the CanESM2 OAnt fingerprint was detectable in all four radiosonde tropospheric temperature datasets. The β values obtained with three of the four radiosonde datasets were close to 2, suggesting that the amplitude of the OAnt fingerprint was too weak relative to observations. In contrast, the CanESM2 AERO fingerprint used by Swart et al. (2018) (15) was not robustly detectable in either ocean temperature or salinity, with mean β values smaller than 1. The OAnt results for surface temperature obtained by Gillett et al. (2013) (27) were similar to the Swart et al. (2018) (15) AERO results. It is unclear why the Lott et al. β value for the CanESM2 OAnt fingerprint differs from both Gillett et al. β value for this fingerprint and the Swart et al. β values for the CanESM2 AERO fingerprint[‖].

In summary, the evidence from the four different climate variables analyzed in Lott et al., Gillett et al., and Swart et al. (15, 26, 27) suggests that the CanESM2 GHG signal may be larger than in observations. This is consistent with the larger than observed tropospheric warming that we show in our Fig. 1. The evidence is more equivocal regarding the question of whether the size of the model-predicted anthropogenic aerosol signal is larger or smaller than in observations.

Finally, we note that CESM1-CAM5 and CanESM2 have very different partitioning of direct and indirect anthropogenic aerosol forcing (29). Neither model provides estimates of the individual (and possibly different) patterns of temperature response to direct and indirect forcing. Such information would be required to assess the credibility of the large indirect aerosol effect in CESM1-CAM5 and the substantially smaller indirect aerosol effect in CanESM2.

References

1. Fu Q, Johanson CM (2005) Satellite-derived vertical dependence of tropical tropospheric temperature trends. *Geophys. Res. Lett.* 32:L10703.
2. Fu Q, Johanson CM (2004) Stratospheric influences on MSU-derived tropospheric temperature trends: A direct error analysis. *J. Clim.* 17:4636–4640.
3. Gillett NP, Santer BD, Weaver AJ (2004) Quantifying the influence of stratospheric cooling on satellite-derived tropospheric temperature trends. *Nature* 432.
4. Kiehl JT, Caron J, Hack JJ (2005) On using global climate model simulations to assess the accuracy of MSU retrieval methods for tropospheric warming trends. *J. Clim.* 18:2533–2539.
5. van den Dool HM, Saha S, Johansson Å (2000) Empirical orthogonal teleconnections. *J. Clim.* 13:1421–1435.
6. Hasselmann K (1979) *On the signal-to-noise problem in atmospheric response studies.* (Roy. Met. Soc., London), pp. 251–259.
7. Santer BD, et al. (2003) Influence of satellite data uncertainties on the detection of externally forced climate change. *Science* 300:1280–1284.
8. Rodgers KB, Lin J, Frölicher TL (2015) Emergence of multiple ocean ecosystem drivers in a large ensemble suite with an earth system model. *Biogeosci.* 12:3301–3320.
9. Keller KM, Joos F, Raible CC (2014) Time of emergence of trends in ocean biogeochemistry. *Biogeosci.* 11:3647–3659.
10. Li J, Thompson DWJ, Barnes EA, Solomon (2017) Quantifying the lead time required for a linear trend to emerge from natural climate variability. *J. Clim.* 30:10179–10191.
11. Fischer EM, Beyerle U, Knutti R (2013) Robust spatially aggregated projections of climate extremes. *Nat. Clim. Change* 3:1033–1038.
12. Lehner F, Deser C, Terray L (2017) Towards a new estimate of "Time of Emergence" of anthropogenic warming: Insights from dynamical adjustment and a large initial-condition model ensemble. *J. Clim.* 30:7739–7756.

[†] The CanESM2 results shown in Lott et al. did not rely on the CanESM2 LEs – they were based on 5 realizations each of the CanESM2 ALL, NAT, and GHG runs (see Table 1 in (26)).

[‡] OAnt has forcing by stratospheric ozone, anthropogenic aerosols, and land use changes.

[§] See Fig. 2 in Swart et al. (2018) (15).

[¶] See Fig. 4a in Gillett et al. (2013) (27) and Fig. 10.4 in Bindoff et al. (2013) (28).

[‖] This difference could be due to multiple factors: 1) OAnt has stratospheric ozone forcing and land use changes, which are not included in AERO; and 2) Lott et al. (2013) did not rely on CanESM2 LEs.

13. Hawkins E, Sutton R (2012) Time of emergence of climate signals. *Geophys. Res. Lett.* 39:L01702.
14. Zhang H, Delworth TL (2018) Robustness of anthropogenically forced decadal precipitation changes projected for the 21st century. *Nat. Comm.* 9:1150.
15. Swart NC, Gille ST, Fyfe JC, Gillett NP (2018) Recent Southern Ocean warming and freshening driven by greenhouse gas emissions and ozone depletion. *Nat. Geosci.* 11:836–841.
16. Solomon S, et al. (2016) Emergence of healing in the Antarctic ozone layer. *Science* 353:269–274.
17. Solomon S, et al. (2017) Mirrored changes in Antarctic ozone and stratospheric temperature in the late 20th versus early 21st centuries. *J. Geophys. Res.* 122:8940–8950.
18. Gleckler PJ, et al. (2006) Krakatoa lives: The effect of volcanic eruptions on ocean heat content and thermal expansion. *Geophys. Res. Lett.* 33:L17702.
19. Maher N, McGregor S, England MH, Gupta AS (2015) Effects of volcanism on tropical variability. *Geophys. Res. Lett.* 42:6024–6033.
20. Pausata FSR, Chafik L, Caballero R, Battisti DS (2015) Impacts of high-latitude volcanic eruptions on ENSO and AMOC. *Proc. Natl. Acad. Sci.* 112:13784–13788.
21. Lee J, Sperber KR, Gleckler PJ, Bonfils CJW, Taylor KE (2019) Quantifying the agreement between observed and simulated extratropical modes of interannual variability. *Cli. Dyn.* 52:4057–4089.
22. Santer BD, et al. (2011) Separating signal and noise in atmospheric temperature changes: The importance of timescale. *J. Geophys. Res.* 116:D22105.
23. Santer BD, et al. (2013) Human and natural influences on the changing thermal structure of the atmosphere. *Proc. Nat. Acad. Sci.* 110:17235–17240.
24. Santer BD, et al. (2018) Human influence on the seasonal cycle of tropospheric temperature. *Science* 361:eaas8806.
25. Allen MR, Tett SFB (1999) Checking for model consistency in optimal fingerprinting. *Cli. Dyn.* 15:419–434.
26. Lott FC, et al. (2013) Models versus radiosondes in the free atmosphere: A new detection and attribution analysis of temperature. *J. Geophys. Res. Atmos.* 118:2609–2619.
27. Gillett NP, Arora VK, Matthews D, Stott PA, Allen MR (2013) Constraining the ratio of global warming to cumulative CO_2 emissions using CMIP5 simulations. *J. Clim.* 26:6844–6858.
28. Bindoff NL, et al. (2013) Detection and Attribution of Climate Change: from Global to Regional 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*, eds. Stocker TF, et al. (Cambridge University Press), pp. 867–952.
29. Zelinka MD, Andrews T, Forster PM, Taylor KE (2014) Quantifying components of aerosol-cloud-radiation interactions in climate models. *J. Geophys. Res.* 119:7599–7615.

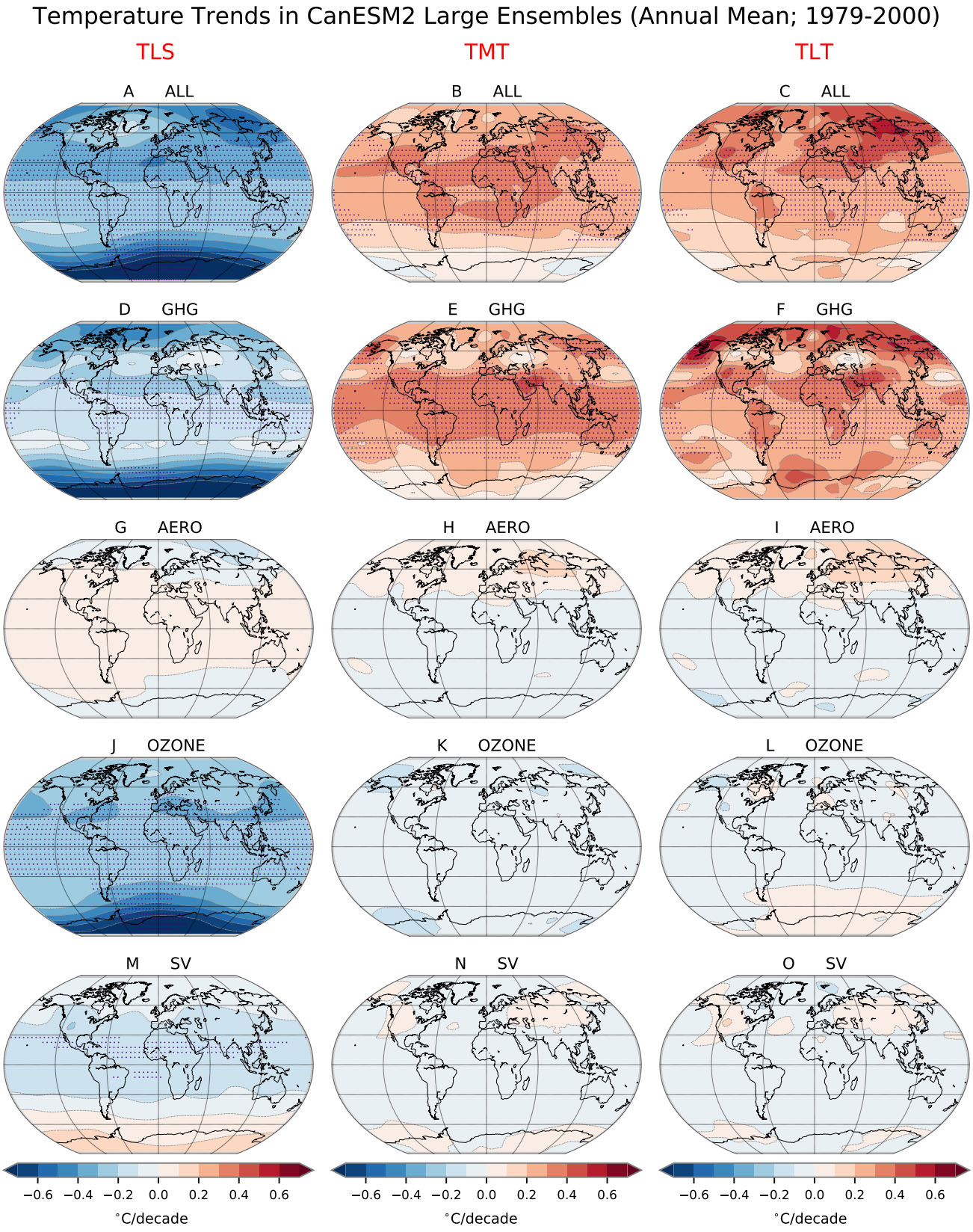

Fig. S1. Ensemble-mean trends in annual-mean temperature over the primary ozone depletion period (1979 to 2000). Results are for the CanESM2 ALL, GHG, AERO, OZONE, and SV ensembles (rows 1-5, respectively) and for TLS, TMT, and TLT (columns 1-3, respectively). The stippling denotes grid-points at which the local S/N ratio is 2 or greater – $i.e.$, grid-points at which the ensemble-mean trend over 1979 to 2000 is at least two times larger than the standard deviation of the 50 individual realizations of the 1979 to 2000 temperature trend. The GHG ensemble is estimated by subtraction.

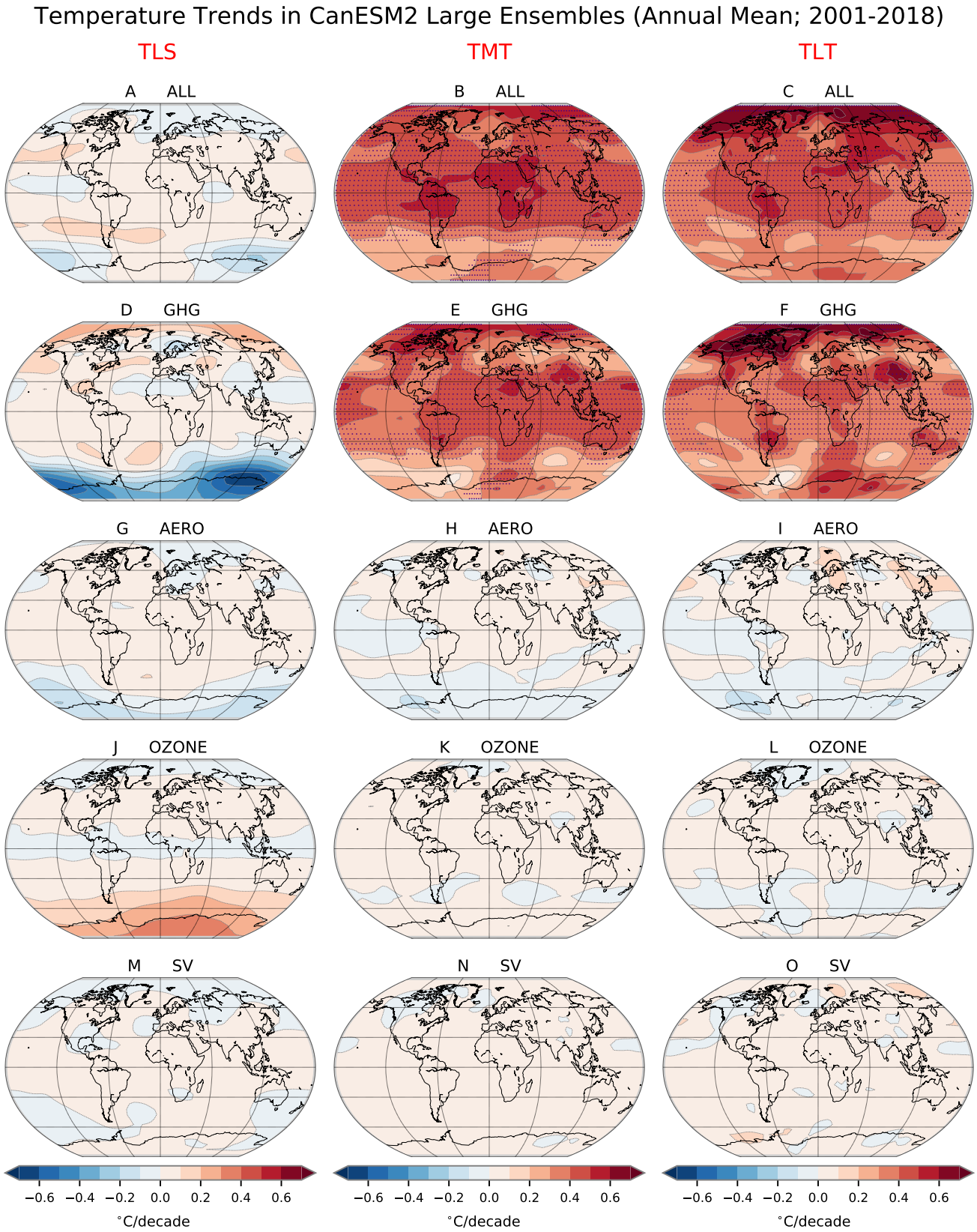

Fig. S2. As for SI Fig. S1 but for trends over the period of partial recovery of lower stratospheric ozone (2001 to 2018).

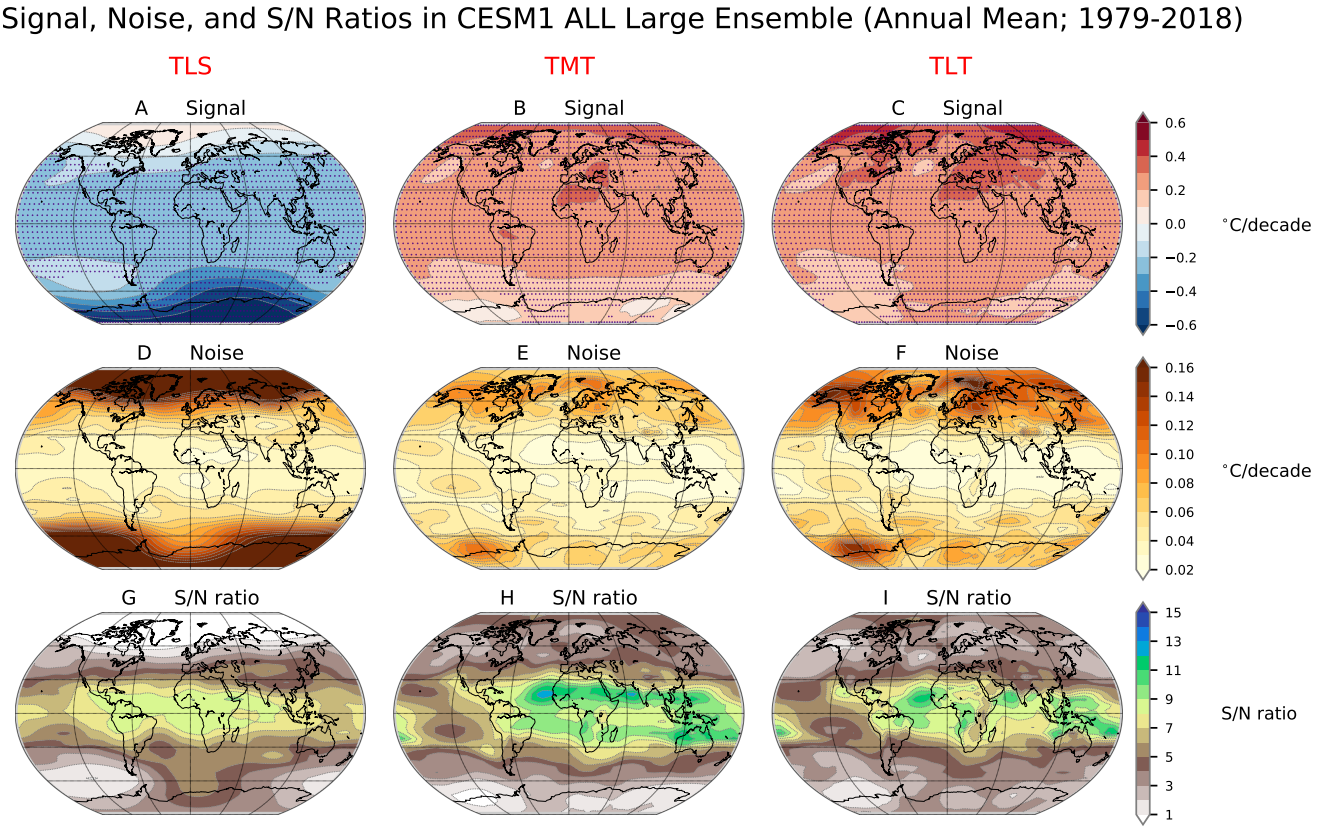

Fig. S3. As for Fig. 3 but for results from the 40-member CESM1 ALL ensemble.

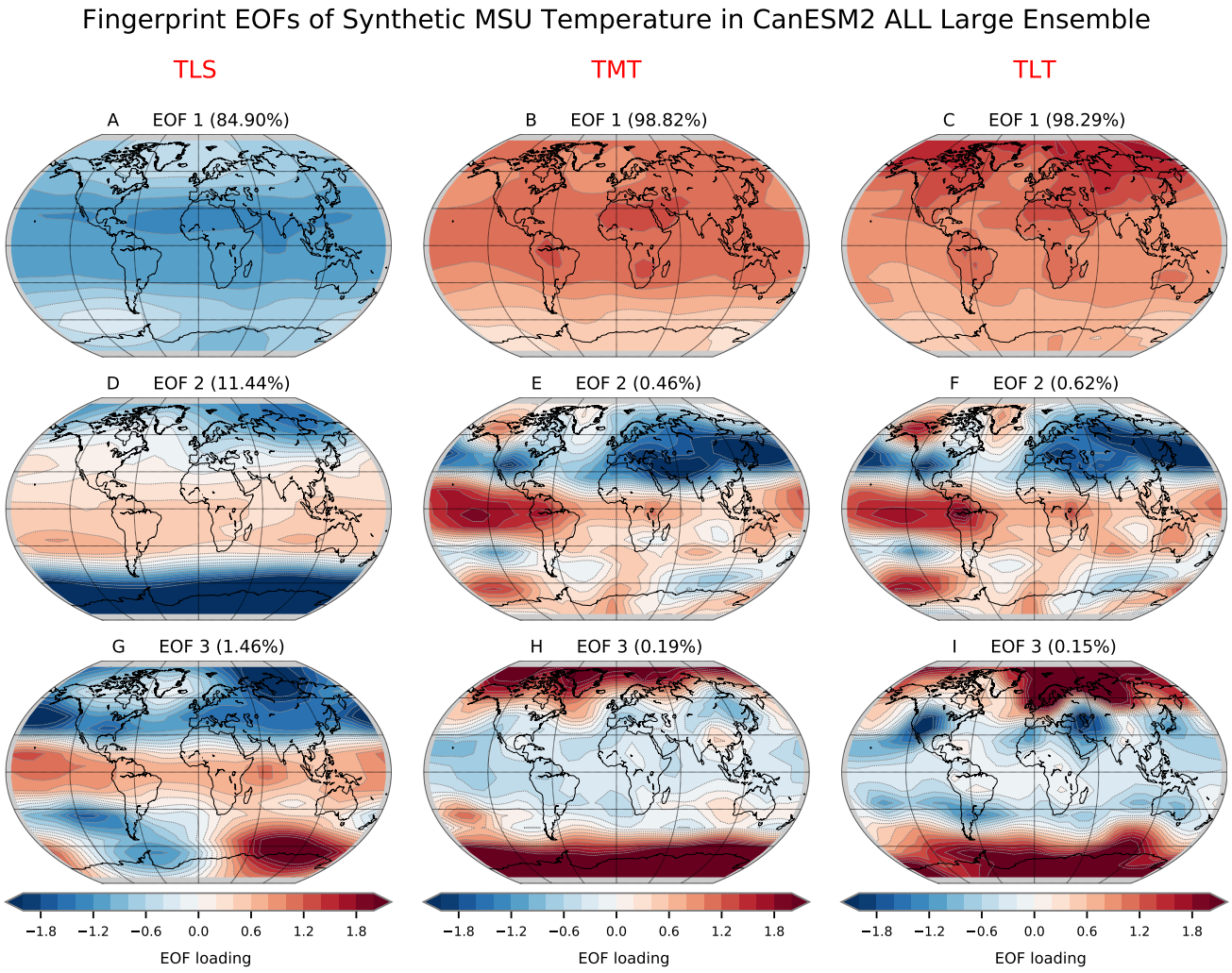

Fig. S4. Leading three EOFs of annual-mean synthetic satellite temperature from the CanESM2 ALL ensemble. Results are for TLS, TMT, and TLT (left, middle, and right columns, respectively). In the pattern-based detection and attribution analysis applied here, the first EOF (row 1) is the fingerprint of the temperature response to combined anthropogenic and natural external forcing (see *Materials and Methods*). EOFs were calculated from the CanESM2 ALL ensemble-average temperature anomalies using simulation output over the satellite era (1979 to 2018). Anomalies were defined relative to climatological annual means for this period.

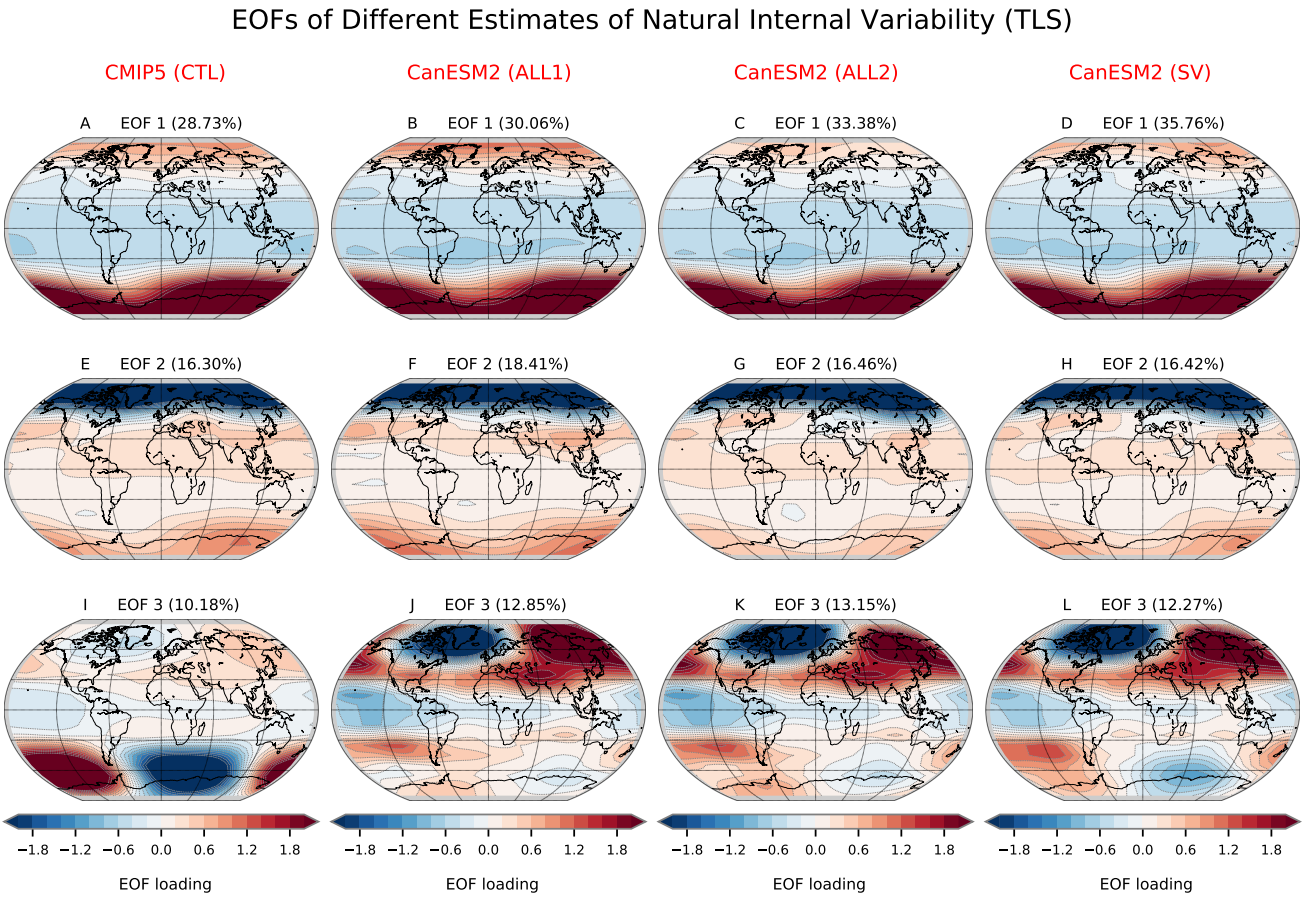

Fig. S5. Leading three EOFs of natural internal variability for annual-mean TLS. Internal variability was estimated from four different sources: 1) CMIP5: the concatenated anomalies from the (detrended) final 200 years of 36 different CMIP5 pre-industrial control runs; 2) ALL1: the concatenated residuals after subtracting the ensemble-mean CanESM2 ALL signal over 1950 to 2100 from the same period of each individual ALL realization; 3) ALL2: the concatenated residuals after subtracting the ensemble-mean CanESM2 ALL signal over 1979 to 2018 from the same period of each individual ALL realization; 4) SV: the concatenated residuals after subtracting the ensemble-mean CanESM2 SV signal over 1950 to 2020 from the same period of each individual SV realization. The sample sizes (in years) for CMIP5, ALL1, ALL2, and SV are 7200 (36 $\times$ 200), 7550 (50 $\times$ 151), 2000 (50 $\times$ 40), and 3550 (50 $\times$ 71). The variance explained is listed above each EOF.

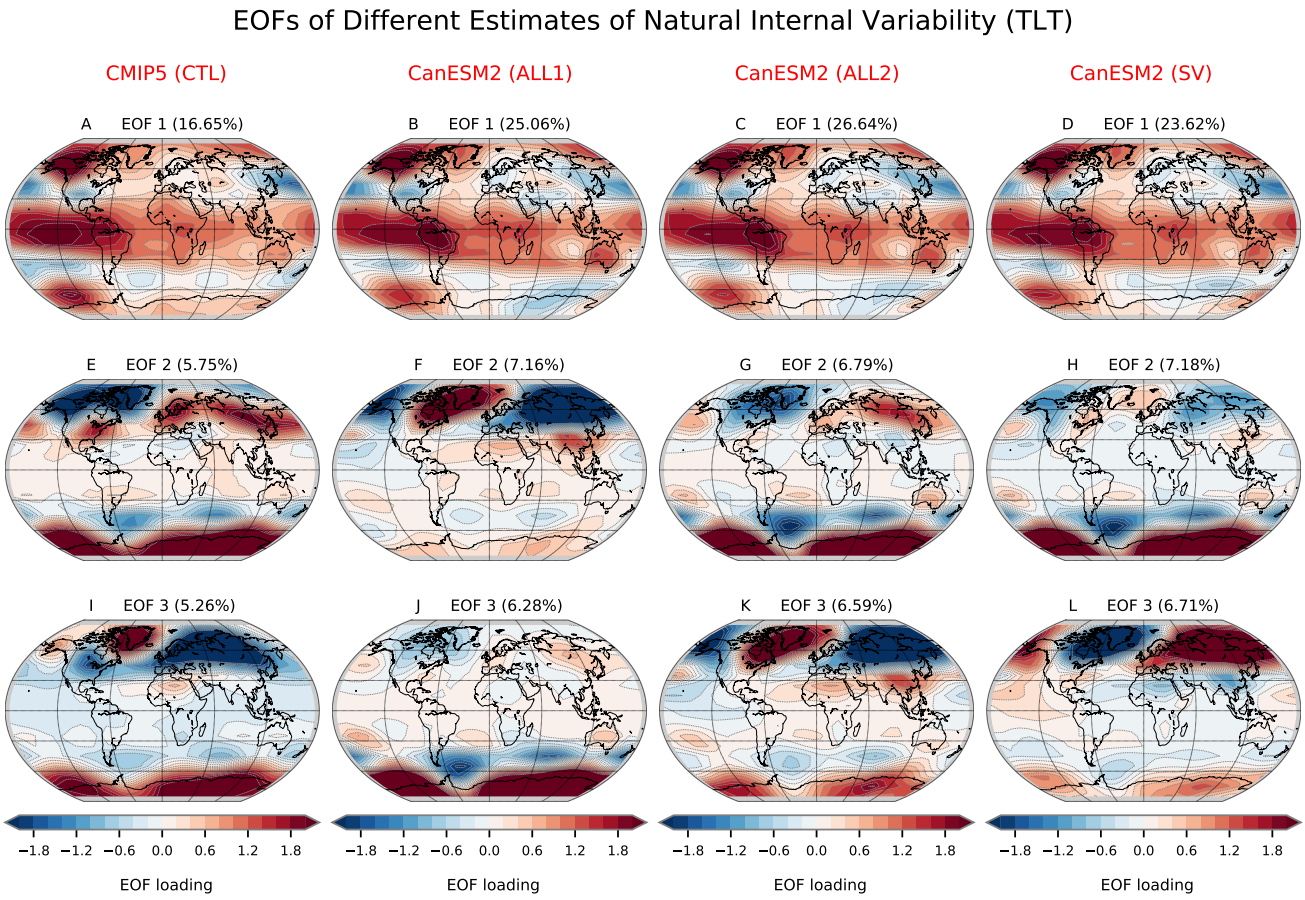

Fig. S6. As for SI Fig. S5 but for annual-mean TLT.

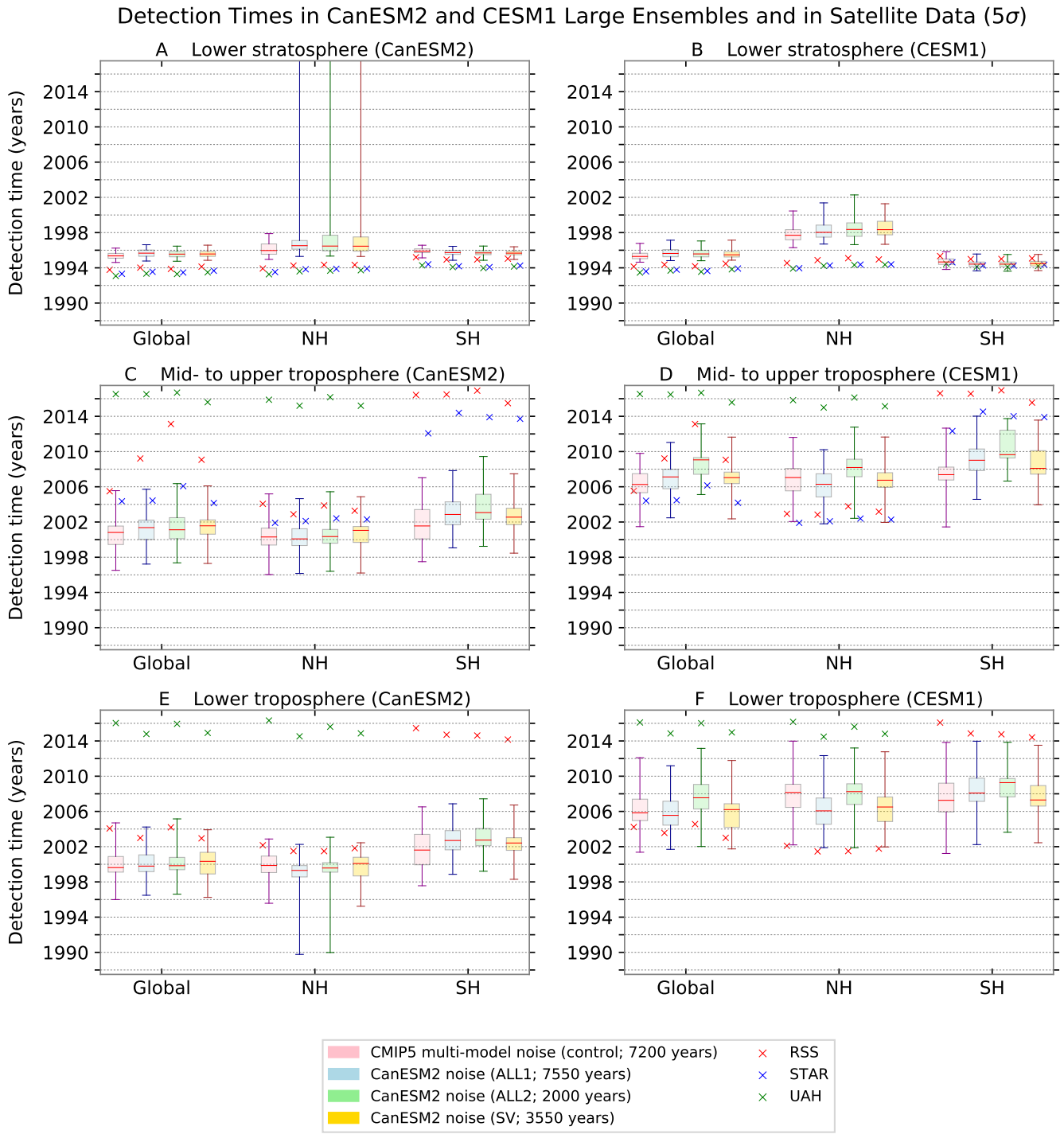

Fig. S7. Fractional fingerprint detection time t_d for TLS (panels A-B), TMT (panels C-D), and TLT (panels E-F). In the left column, fingerprints estimated from the CanESM2 ALL simulation are searched for in satellite data sets and in each of the 50 individual CanESM2 ALL realizations. In the right column, CESM1 fingerprints are compared with satellite data and with each of 40 individual CESM1 ALL realizations. Results are as for Figure 5 in the main text, but detection times were calculated with a more stringent 5σ significance threshold. For TLS changes for the NH domain, three of the four CanESM2 ALL detection time distributions have very large $t_{d\{r\}}$ ranges (panel A). In these three cases, there are a small number of ALL realizations in which the TLS fingerprint is not detectable until late in the 21st century. By chance, the S/N ratios in these anomalous realizations are very close to the 5σ threshold, and dip above and below the threshold as the analysis period increases. This leads to a bifurcation in the distribution of detection times, and explains why the "whiskers" extend to the end of the 21st century even though $t_{d\{m\}}$ is in 1996.