Supplemental Material

Supplemental Material: Model Selection Procedure

This Supplemental section gives a detailed description of the procedure used to select CMIP5 models or model groupings suitable for Southern Ocean Temperature and Salinity change detection analysis. The criteria for selection were:

1) a sufficiently long control simulation to represent internal variability; for these study, we chose a minimum of 400 years, being 10 realizations (i.e. independent 40-year periods).
2) a minimum of 3 members of the 'historical' simulation. Ideally more than 3 members are required to average out noise from the estimated response to forcings, but more than 3 would be very restrictive model selection in the context of available output in the CMIP5 archive.
3) a reasonable representation of Southern Ocean watermasses in the mean state

23 models satisfied the first 2 criteria, which were then analyzed with respect to their representation of the 4 major Southern Ocean watermasses: Sub-Antarctic Mode Water (SAMW). Antarctic Intermediate Water (AAIW), Circumpolar Deep Water (CDW) and Antarctic Bottom Water (AABW). Many CMIP5 models have a poor representation of these masses (Sallee *et al*, 2013), and since their formation is critical for Southern Ocean heat uptake, we argue that some model selection is justified for confidence in our Detection and Attribution.

We used the method of (Roy *et al*, submitted) to estimate the potential density boundaries, applied to the 1965-2005 mean state of one 'historical' ensemble member from each of the 23 models. This method uses climatological profiles of salinity (S) and Potential Vorticity (PV) at 30^{o}S, mapped onto density coordinates, to identify the density ranges of AAIW (defined by a pronounced salinity minimum at 30^{O}S) and SAMW (defined by a local, shallow minimum of PV). AABW is defined by a strong, deep PV minimum in the zonal-mean PV from the Antarctic coast to 30^{o}S; CDW is the residual watermass between AAIW and AABW. These boundaries are presented in CSV format in Table S1.

Having defined the density boundaries, the 1966-2005 time-mean temperature and salinity was calculated for each model's watermasses (i.e. the means for volume south of 30^{o}S bounded by the density boundaries). We compared these mean properties – which are given in Table S2 and plotted in Figure S1 - with watermass T and S estimated for the CARS climatology (Ridgway *et al*, 2002) using the same objective method. Models were excluded from the study if they demonstrated noticeable deviation in any of the watermasses compared to the CARS climatology and the multi-model distribution. This selection was done by visual inspection of the T-S diagrams rather than using a quantitative metric of skill; this does mean that the selection is subjective, but we argue that the definition of a quantitative threshold for selection or rejection is itself somewhat arbitrary. We note that the distribution of the accepted models states in Figure S1 is clearly in closer agreement to the CARS climatology than the distribution of the rejected models, with the exception of CanESM2; this model was retained, despite its relatively poor representation of SAMW, AAIW and CDW, for direct comparison with Swart et al, (2018).

We found that the ACCESS models (ACCESS1.3 and ACCESS1.0), which have similar ocean and atmosphere dynamic cores, have very similar watermass properties, therefore there were merged to create a single model ensemble (ACCESS1) with more ensemble members and piControl realisations (hence greater statistical robustness). We also found models that did not satisfy the piControl length and ensemble member criteria, but which were similar in their model code and watermass states to selected models; this was the case for MIROC-ESM-CHEM (related to MIROC-ESM) and MPI-ESM-P (related to MPI-ESM-LR); as with the ACCESS model, these were merged with their related model to increase piControl realisations and the size of the forced ensemble.

In general, intermodal spread is largest in the shallower watermasses (SAMW and AAIW). There is a fresh SAMW bias in almost all the models and a wide spread of salinities, even amongst the selected models. However, the isopycnals in Figure S1 suggest that temperature is at least as important as salinity in setting the density bias of SAMW. The selected models have a reasonable volume of SAMW compared to the CRAS climatology, but the rejected models have quite large volume biases, both positive and negative.

For AAIW there is no clear multimodel salinity bias in the selected models, although they tend to be warmer and lighter than the climatology. With the exception of a few outliers, the models have similar density in the denser watermasses (CDW and AABW) as the climatology, with a range of temperatures, but with salinity apparently a stronger constraint on watermass density. None of the models have noticeably larger volumes of CDW or AABW than the climatology, although several have very small volumes of AABW.

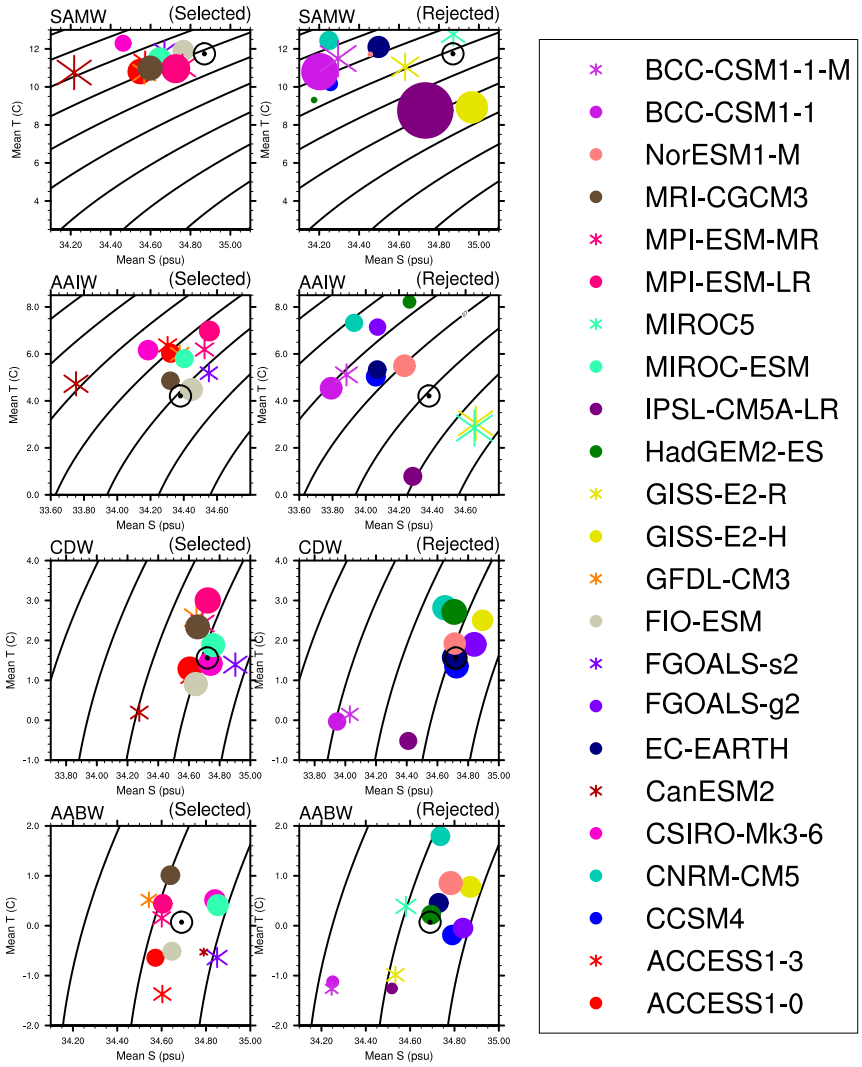

Figure S1 – 1965-2005 watermass mean T and S, for each model and watermass; black circled dot shows the estimate from the CARS climatology. Contour lines show 2000m potential density contours. The 13 models selected for use in this study are shown in the left hand column; 'rejected' models are shown in the right hand column. The size of each marker is scaled by the volume of that watermass in each model

```
model,expt,SAMW_upper,SAMW_IW, IW_CDW,CDW_BW
ACCESS1-0,historical,34.753,35.334,36.476,37.236
ACCESS1-3,historical,34.693,35.254,36.436,37.324
CCSM4,historical,33.851,35.244,36.546,37.351
CNRM-CM5,historical,33.711,34.943,36.095,37.051
CSIRO-Mk3-6,historical,34.723,35.314,36.546,37.317
CanESM2,historical,34.142,35.334,36.396,37.384
EC-EARTH,historical,34.102,35.404,36.506,37.224
FGOALS-g2,historical,33.881,34.993,36.295,37.364
FGOALS-s2,historical,34.172,35.604,36.836,37.429
FIO-ESM,historical,34.322,35.694,36.856,37.279
GFDL-CM3,historical,34.142,35.404,36.466,37.076
GISS-E2-H,historical,34.282,35.955,36.997,37.224
GISS-E2-R,historical,34.172,35.905,37.277,37.212
HadGEM2-ES,historical,34.032,34.843,35.905,37.223
IPSL-CM5B-LR,historical,34.543,36.125,37.007,37.137
MIROC-ESM,historical,34.312,35.464,36.596,37.298
MIROC-ESM-CHEM,historical,34.282,35.434,36.586,37.320
MIROC5,historical,34.242,35.584,37.147,37.117
MPI-ESM-LR,historical,34.372,35.444,36.406,37.121
MPI-ESM-P,historical,34.362,35.474,36.416,37.133
MPI-ESM-MR,historical,34.372,35.534,36.526,37.149
MRI-CGCM3,historical,34.022,35.494,36.616,37.080
NorESM1-M,historical,35.093,35.214,36.646,37.223
bcc-csm1-1,historical,34.472,35.284,36.336,36.995
bcc-csm1-1-m,historical,34.412,35.254,36.376,37.003
```

Table S1 – 2000m potential density boundaries for each model's historical simulation Southern Ocean watermasses south of 30°S, in CSV format. The boundaries are the upper SAMW (SAMW_upper), SAMW-AAIW interface (SAMW_IW), AAIW-CDW interface (IW_CDW) and CDW-AABW interface (CDW_BW).

Model	SAMW			AAIW			CDW			AABW		
	ϑ (°C)	S (g/kg)	V (10^{17} m^3)	ϑ (°C)	S (g/kg)	V (10^{17} m^3)	ϑ (°C)	S (g/kg)	V (10^{17} m^3)	ϑ (°C)	S (g/kg)	V (10^{17} m^3)
CARS climatology	11.74	34.87	0.23	4.21	34.38	0.89	1.56	34.72	2.15	0.07	34.69	1.20
Selected models												
ACCESS1-0	10.81	34.55	0.31	6.03	34.32	0.48	1.29	34.61	2.66	-0.64	34.57	0.52
ACCESS1-3	11.29	34.57	0.26	6.36	34.3	0.49	1.13	34.62	2.66	-1.37	34.6	0.58
CSIRO-Mk3-6	12.30	34.46	0.08	6.16	34.18	0.60	1.42	34.74	2.62	0.52	34.84	0.94
FIO-ESM	11.91	34.76	0.18	4.48	34.45	0.88	0.91	34.64	2.50	-0.51	34.65	0.62
FGOALS-s2	11.86	34.67	0.22	5.18	34.55	0.43	1.39	34.90	2.66	-0.64	34.85	0.91
GFDL-CM3	10.75	34.57	0.27	5.99	34.38	0.55	2.57	34.65	3.00	0.52	34.54	0.41
MIROC-ESM	11.60	34.64	0.24	5.88	34.38	0.45	1.79	34.76	2.50	0.27	34.86	1.01
MIROC-ESM-CHEM*	11.46	34.65	0.25	5.80	34.40	0.46	1.88	34.76	2.46	0.41	34.85	1.02
MPI-ESM-LR	10.96	34.73	0.48	6.98	34.55	0.63	3.00	34.72	3.23	0.44	34.61	0.69
MPI-ESM-P*	10.87	34.76	0.51	6.82	34.57	0.64	2.86	34.71	3.21	0.29	34.60	0.67
MPI-ESM-MR	11.03	34.76	0.41	6.19	34.52	0.52	2.42	34.69	2.55	0.15	34.60	0.63
MRI-CGCM3	10.99	34.60	0.30	4.86	34.32	0.42	2.34	34.66	2.71	1.01	34.64	0.71
CanESM2	10.75	34.22	0.91	4.73	33.75	1.18	0.19	34.28	1.10	-0.53	34.79	0.05
Rejected models												
BCC-CSM1-1	11.49	34.29	1.06	5.11	33.88	0.90	0.15	34.03	0.97	-1.26	34.25	0.23
BCC-CSM1-1-M	10.8	34.20	1.06	4.54	33.79	0.83	-0.03	33.95	0.99	-1.13	34.25	0.16
CCSM4	10.16	34.26	0.05	5.04	34.06	0.54	1.35	34.72	2.63	-0.19	34.79	0.80
CNRM-CM5	10.72	34.44	0.16	5.84	34.24	0.37	1.83	34.70	3.13	-0.34	34.65	0.42
EC-EARTH	12.10	34.50	0.20	5.33	34.07	0.44	1.57	34.71	2.87	0.46	34.73	0.74
FGOALS-g2	13.25	34.34	0.11	7.14	34.07	0.34	1.90	34.84	2.68	-0.05	34.84	0.81
GISS-E2-H	14.23	34.54	0.15	5.80	34.18	0.87	1.71	34.52	2.12	-0.02	34.64	1.02
GISS-E2-R	10.60	34.63	0.52	3.22	34.66	3.28	-	-	0.00	-1.03	34.52	0.58
HadGEM2-ES	9.31	34.17	0.00	8.22	34.26	0.15	2.71	34.71	3.09	0.22	34.69	0.72
IPSL-CM5A-LR	8.76	34.73	3.90	0.79	34.28	0.46	-0.52	34.41	0.93	-1.26	34.52	0.12
MIROC5	12.77	34.87	0.28	2.84	34.65	4.14	-	-	0.00	0.39	34.58	0.96
NorESM1-M	11.71	34.46	0.00	5.50	34.23	0.82	1.92	34.71	1.95	0.85	34.78	1.37

*Table S2 1965-2005 historical mean watermass potential temperature, salinity and volume (i.e. values plotted in Figure S1), along with reference values from CARS climatology. * indicates models which did not satisfy the piControl length or ensemble size criteria, but were sufficiently similar to a selected model to be merged with that model. CanESM2 has watermass properties closer to the rejected set than the accepted set, but was included for direct comparison with Swart et al (2018)*

Further Information on Optimal Fingerprinting Procedure

In this section, we give further information on how the appropriate truncation level, i.e. number of retained Empirical Orthogonal Functions (EOFs) was selected. As an example, we take the temperature Attribution results from the ACCESS1 ensemble (i.e. the merging of ACCESS1-0 and ACCESS1-3).

The results are summarized in Figure S2, where k is the truncation level. In the upper panel, the scaling factors for each response - GHG-only (red), Natural-only (green) and residual (blue) - is shown by markers, with vertical lines showing the range due to internal variability alone. Recall that a response is 'detectable' where its range does not include zero. Figure S2a shows that as more EOFs are retained the range due to internal variability tends to decrease, which is typical but means that there is a dependency on the results (Allen and Tett, 1999). As a first pass, we apply a consistency check as described by Allen and Tett (1999). Assuming that the observed change is the linear sum of all scaled responses plus internal variability (and that the model is reasonable in its representation of internal variability), then the residual of the observations minus the scaled responses should be approximately equal to the simulated internal variability, i.e.

$$r2 = \left(Y - \sum \beta_i X_i\right) \boldsymbol{C_n}^{-1} \sim 1$$

Where r2 is the test statistic, Y is the observed pattern, X_i are the individual response, β_i are the scaling factors, and $\mathbf{C_n^{-1}}$ is the inverse noise covariance matrix (from the model piControl simulation). This test is summarized in Figure S2b, where the dashed curves show the 99% confidence interval that this ratio is 1, based on a chi-squared distribution with k-Nexpt degrees-of-freedom, where Nexpt is the number of response patterns (i.e. 3).

Truncation levels of $16 \geq k \geq 10$ pass this test (i.e. the ratio is within the 99%-confidence intervals. However, even within this range there is some dependency; the GHG signal is detectable for $13 \geq k$, and the residual (i.e. ozone plus anthropogenic aerosols) response is detectable only for k = 12 and 13. Recognizing that the aim of the optimal fingerprint technique is maximize 'detectability', we choose the maximum truncation level that a) passes the consistency test, and b) gives the maximum number of detectable signals. In this case, k = 12-13 passes these tests (with 2 detectable responses, GHG-only and the ozone-plus-aerosol response), and so k = 13 is selected as the maximum of these truncation levels.

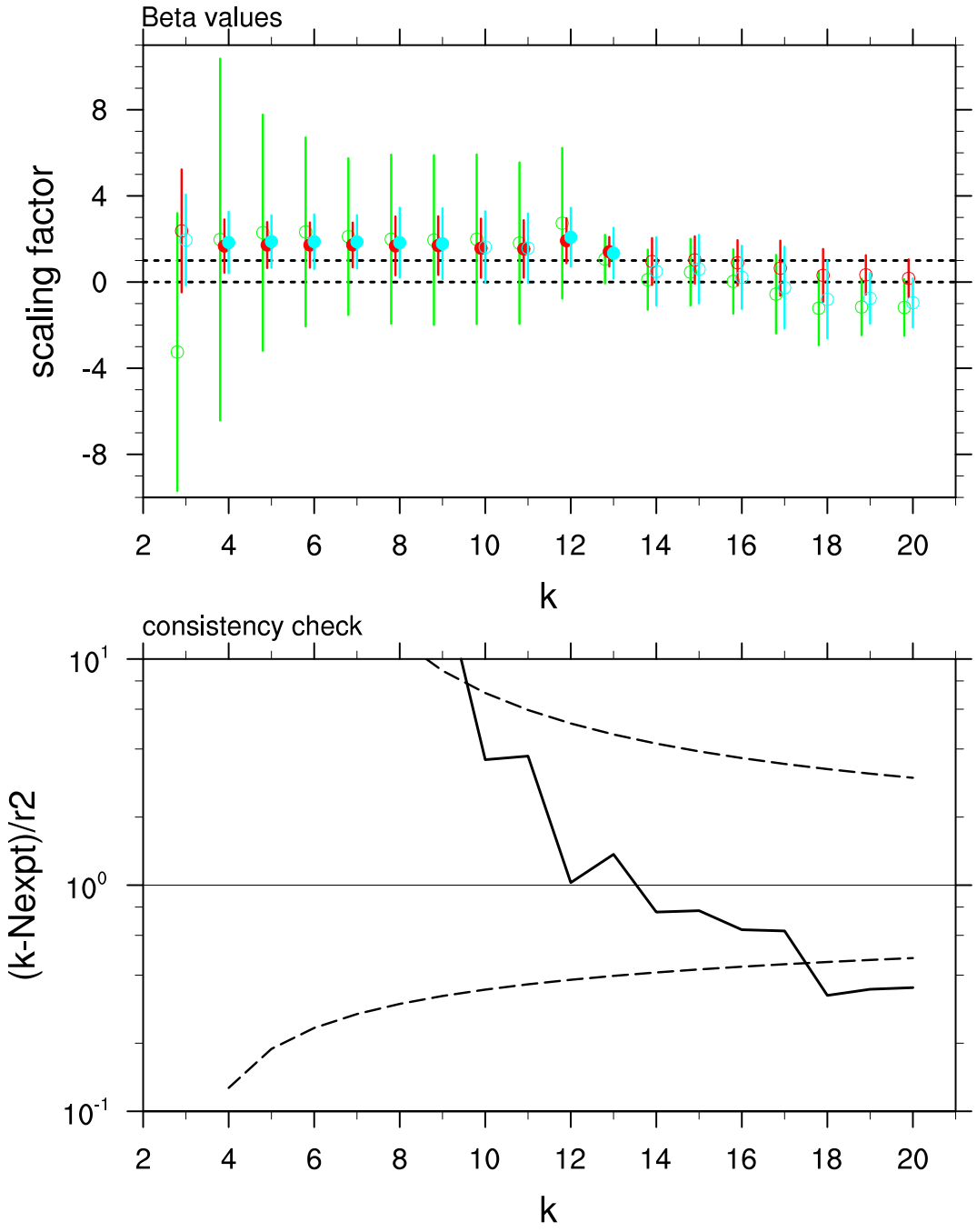

Figure S2 – Results of temperature optimal fingerprint analysis for ACCESS1 model ensemble. Upper panel shows best fit scaling factors (markers, y-axis) by truncation level (k, x-axis), for GHG-only (red), Natural-only (green) and aerosol-plus-ozone responses (blue). Vertical lines show the range due to internal-variability; when this reange does not include 0 the response is detectable (shown by solid marker); non-detectable ranges are shown by open markers. Bottom panel shows the results of the Allen and Tett (1999) consistency test, where dashed lines show the 99% confidence limits estimated using a chi-squared distribution with k-Nexpt degrees-of-freedom; results are consistent with the model internal variability when they are within the confidence interval.

References

Allen, M. R., and S. F. B. Tett, 1999: Checking for model consistency in optimal fingerprinting. *Climate Dynamics*, **15,** 419-434, Doi 10.1007/S003820050291.

Ridgway, K. R., J. R. Dunn, and J. L. Wilkin, 2002: Ocean interpolation by four-dimensional least squares -Application to the waters around Australia. *J Atmos Ocean Tech*, **19,** 1357.

Roy, T., J. B. Sallee, L. Bopp, and N. Metzl, submitted: Diagnosing the drivers of Southern Ocean climate-carbon cycle feedbacks in multiple Earth system models using a semi-automated water mass tracking approach.

Sallee, J. B., E. Shuckburgh, N. Bruneau, A. J. S. Meijers, T. J. Bracegirdle, Z. Wang, and T. Roy, 2013: Assessment of Southern Ocean water mass circulation and characteristics in CMIP5 models: Historical bias and forcing response. *J Geophys Res-Oceans*, **118,** 1830-1844, 10.1002/jgrc.20135.

Swart, N. C., S. T. Gille, J. C. Fyfe, and N. P. Gillett, 2018: Recent Southern Ocean warming and freshening driven by greenhouse gas emissions and ozone depletion. *Nat Geosci*, 10.1038/s41561-018-0226-1.