JGR Oceans

Key Points:
- We perform a comprehensive sensitivity study of the heat distribution in mode water formation regions
- Sensitivities are highest to local same-winter heat flux changes, and both local and remote wind stress changes up to at least 8 years past
- High sensitivity regions reveal kinematic links with source waters and dynamic links with boundary current regions

Correspondence to:
E. J. D. Boland,
emmomp@bas.ac.uk

Local and Remote Influences on the Heat Content of Southern Ocean Mode Water Formation Regions

Emma J. D. Boland[1] , Daniel C. Jones[1] , Andrew J. S. Meijers[1] , Gael Forget[2] , and Simon A. Josey[3]

[1]British Antarctic Survey, Cambridge, UK, [2]Massachusetts Institute of Technology, Cambridge, MA, USA, [3]National Oceanography Centre, Southampton, UK

Abstract The Southern Ocean (SO) is a crucial region for the global ocean uptake of heat and carbon. There are large uncertainties in the observations of fluxes of heat and carbon between the atmosphere and the ocean mixed layer, which lead to large uncertainties in the amount entering into the global overturning circulation. In order to better understand where and when fluxes of heat and momentum have the largest impact on near-surface heat content, we use an adjoint model to calculate the linear sensitivities of heat content in SO mode water formation regions (MWFRs) to surface fluxes. We find that the heat content of these regions is, in all three basins, most sensitive to same-winter, local heat fluxes, and to local and remote wind one to eight years (the maximum lead-time of our simulations) previously. This is supported by sensitivities to potential temperature changes, which reveal the sources of the MWFRs as well as dynamic links with boundary current regions and the Antarctic Circumpolar Current. We use the adjoint sensitivity fields to design a set of targeted perturbation experiments, allowing us to examine the linear and non-linear responses of the heat content to changes in surface forcing. In these targeted experiments, the heat content is sensitive to both temperature changes and mixed layer volume changes in roughly equal magnitude.

Plain Language Summary The Southern Ocean (SO) is of crucial importance to the global ocean's uptake of carbon and heat. However, due to difficulties in making observations in such a remote and hostile environment, we currently don't know accurately how much heat and carbon enters the SO from the atmosphere. Heat from the SO can get locked away for hundreds to thousands of years in the world's deep oceans, entering through a few key regions. We use a computer model to assess how the heat, fresh water, and wind energy entering through the surface of the SO affects the heat of these key regions. We find that these regions are very sensitive to heat coming in through the surface directly over them, and that winds across a wider area of the SO can affect the heat stored for several years. If we want to estimate the heat stored in these regions more accurately, this information can be used to help us decide where and when it is important to measure the winds and heat entering the ocean better.

1. Background

The Southern Ocean (SO) is home to the world's longest and strongest ocean current, the Antarctic Circumpolar Current (ACC), which encircles the globe free of continental barriers. Driven by strong wind and buoyancy forcing, the ACC transports climatically important tracers such as heat, salinity, and carbon between the three major ocean basins. These forcings also create sloping density surfaces (isopycnals) that tilt upwards from north to south, which connect deep waters from around the globe to the surface. At the surface, air-sea interactions modify the properties of water masses. These modified waters then return to depth and into the other ocean basins as dense waters near the Antarctic continental shelf, or as lighter mode and intermediate waters north of the ACC (Lumpkin & Speer, 2007; Marshall & Speer, 2012).

The SO is of critical importance to the global oceanic uptake of heat and carbon, due in part to this overturning circulation. It may be responsible for as much as 75% of the global ocean heat uptake and ~50% of the carbon uptake (Frölicher et al., 2015; Mikaloff Fletcher et al., 2006). Roughly 30% of anthropogenic CO_2 emissions ends up in the ocean (Khatiwala et al., 2013), and around 93% of the excess heat added to the earth system since 1955 has been estimated to be stored in the ocean (Levitus et al., 2012), predominantly in the SO (Roemmich et al., 2015).

Understanding what determines the time scales of SO overturning and the properties of the waters transported is of crucial importance to future climate predictions, including the continued efficiency of the carbon sink (Landschützer et al., 2015; Le Quéré et al., 2018). The properties of the overturning circulation are affected by a range of processes, including variations in surface forcings, variations in the interactions of these forcings with ocean mixed layer properties, and variations in the draw-down of mixed layer properties into the ocean interior as mode, intermediate, and deep waters. Unfortunately, direct air-sea flux observations are scarce in the SO, especially in the winter when sea ice hinders access to the region (Newman et al., 2015).

This work focuses on understanding how variations in surface forcings impact mixed layers in SO mode water formation regions (MWFRs), using a data-constrained estimate of these processes (ECCOv4, Forget, Campin, et al., 2015). This will provide insights into the influence of uncertainties in observations of surface forcings on estimates of mode water properties, as well as allowing for estimates of the impact of future changes in these forcings.

To this end, we follow the same approach as in Jones et al. (2018), using an adjoint model, but here we target the SO mixed layer. The adjoint experiments designed in Section 2 derive the linear sensitivities of MWFRs in the Atlantic, Pacific, and Indian sectors of the SO to surface heat and winds (Section 3). We then decompose the potential temperature sensitivities of these regions into kinematic and dynamic sensitivities (Section 4). The linear sensitivities from the adjoint are then used to design targeted perturbation experiments using the non-linear forward model (Section 5). We finish with a summary of our results, discussion and perspectives (Section 6).

2. Experiment Design

2.1. The ECCOv4 Global State Estimate and Its Adjoint

For this study we used the ECCOv4 (release 2) ocean state estimate framework (Forget, Campin, et al., 2015). This is a global ~1° ocean and sea ice setup of the MITgcm (Adcroft et al., 2004) that spans 20 years from 1992 to 2011, with surface forcings and initial conditions that have been optimized to reduce misfits to observations. Details of the 4D-Var optimization process and the residual model-data misfit can be found in Forget, Campin, et al. (2015). This set-up provides a recent, well-constrained estimate of the SO, which is also easily modified to carry out adjoint sensitivity experiments. We selected ECCOv4 in order to allow for possible dynamical connections with the global ocean, and to work with a two-decade time period. It would be instructive to repeat our experiments in the higher-resolution Southern Ocean State Estimate (SOSE, Mazloff et al., 2010) although SOSE covers a shorter time period and has a boundary in the subtropics.

Mixed layer depths in ECCOv4, used to define mode water formation regions, closely match observations in terms of geography and magnitude (see Figure 6, Forget, Ferreira, & Liang, 2015). Figure S1 in Jones et al. (2019) additionally shows a comparison of the sea level anomaly and sea surface temperatures in ECCOv4 with observations in the Indian and Pacific mixed layer regions also used in this study, showing that absolute values and variability are well captured. ARGO data is used to produce ECCOv4, and ECCOv4 temperature and salinity compare well with ARGO-derived values in the Pacific, Atlantic, and Indian mode water formation regions (Figure 1, S1, and S2). To allow for direct comparison with the monthly ECCOv4 solution, the ARGO profiles in the region from that month, and the months either side, were linearly interpolated to standard depths. The ECCOv4 solution was then subsampled identically (via linear interpolation) to produce a complementary set of profiles for the same three month period, which was then averaged to produce the red lines in Figure 1. The black lines were calculated by taking the sum of the ECCOv4 profile means (red lines) and the median model-data misfit for each three-month period. The model solution shows good general agreement with ARGO for both quantities at all depths, although note the differences near 400 m depth appear larger due to the smaller *y*-axis range.

An adjoint model, in this context, is one that starts from a quantity of interest (henceforth referred to as an "objective function")—such as the integrated temperature over a certain region (e.g., Jones et al., 2018, 2019, 2020), the heat or volume transport of a particular current (e.g., Czeschel, et al, 2010, 2012; Heimbach et al., 2011; Kostov, et al, 2019; Pillar, et al, 2016; Smith & Heimbach, 2019)—and steps backwards through a linearized version of the model, propagating the sensitivities of the objective function.
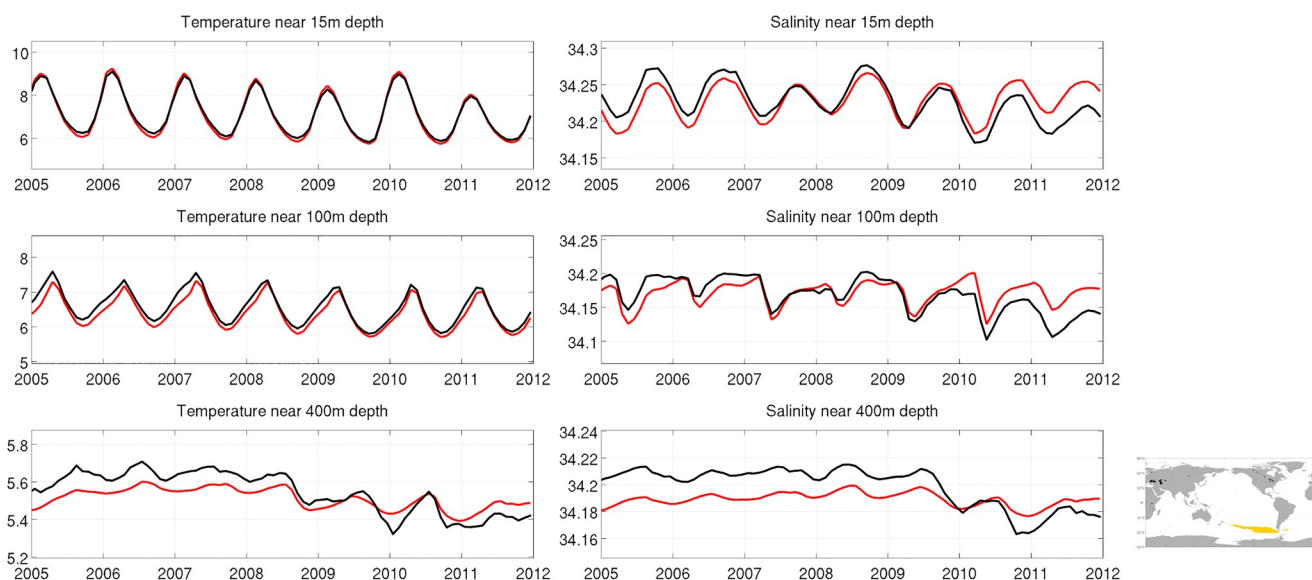
Figure 1. Comparison of direct measurements from ARGO floats (black line, see www.argo.ucsd.edu for more info) and the ECCOv4r2 solution, sub-sampled identically (red lines) with, for potential temperature (left) and salinity (right), in the median Pacific mode water formation region (yellow-shaded area bottom right, see text for how this region is defined). See text for details on the calculation. Note the different *y*-axis scales make the differences near 400 m appear larger, although they are of similar magnitude to the shallower depths.

More detailed descriptions of how sensitivity information propagates backwards through an adjoint model can be found in Heimbach et al. (2011); Marotzke et al. (1999). The adjoint model produces the linear sensitivity of the objective function to a range of specified model variables, such as surface fluxes or interior properties (e.g., potential temperature, mixing parameters). In a more traditional model study, one might start by choosing a model variable or variables theorized to impact one's objective function, and then carry out a suite of perturbation experiments changing these variables by a range of magnitudes, locations, and/or times. In contrast, an adjoint model can produce in one single model the linear sensitivity of one's objective function to a range of model variables, at all points in the model domain, at multiple time lags, allowing for a fully comprehensive experiment.

2.2. Defining the Objective Function: Locating Mode Water Formation Regions

For this study, our objective function was the heat content of the mixed layer in SO mode water formation regions. Mode water is formed seasonally in the deep winter mixed layers to the north of the ACC, before subduction into the interior across the base of the mixed layer (Sallée, et al., 2010; Talley, 1999). By definition, such water is characterized by low stratification (i.e., low potential vorticity [PV] values) (see e.g., Hanawa & Talley, 2001). Figure 2 shows a latitude-depth plot along 90°E (in the Indian sector of the SO) of the minimum PV values for a representative year (1999) from the ECCOv4 r2 state estimate (notice the logarithmic color scale). There is a sharp lateral gradient in the minimum PV values just inside the winter mixed layer extent, and as such the winter mixed layer extent captures the mode water formation pools of interest.

2.2.1. Atlantic, Pacific, and Indian Mode Water Pools

Three distinct mode water formation pools can be identified in the three main basins—Atlantic, Pacific, and Indian (Figure 3a). The winter mixed layer encloses the mode water formation pools (see also Figure 2). We used a combination of annual minimum potential vorticity (PV) values and winter (ASO) mixed layer depths to form the horizontal mask for the "objective function" volume for the suite of adjoint experiments we carried out, whilst ensuring that nothing too close to land or too far north was included. Specifically, we defined the objective function as anywhere between 30°S and 65°S with a minimum PV value of less than 10^{-13} and an ASO mean mixed-layer depth (MLD) (for that given year) of greater than 300 m depth, then manually removed regions in the North of the basins (we removed regions north of 40°S in the Pacific and East Indian Ocean [60°W to 110°E], north of 35°S in the West Indian Ocean [110°E to 60°E] and north of
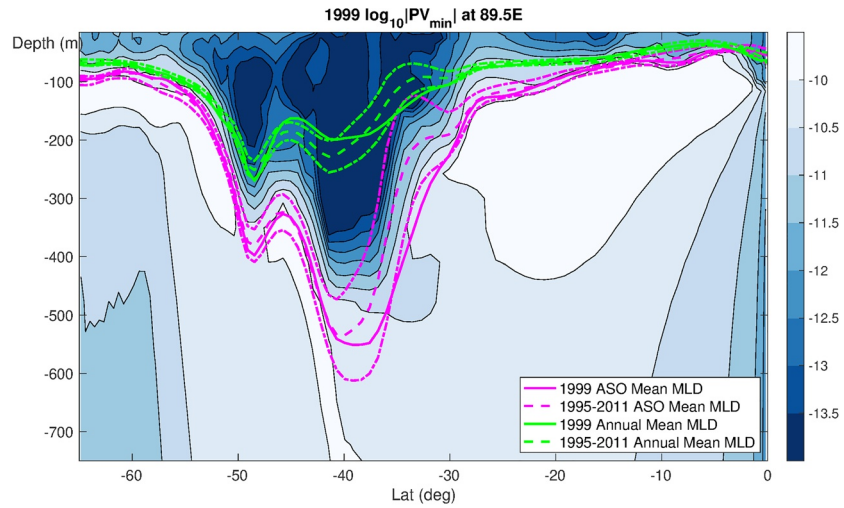
Figure 2. An example mode water formation region, characterized by low potential vorticity (PV) values, contained within the winter mixed layer: Latitude-depth plot of the absolute value of the 1999 minimum PV along 90°E in ECCOv4 r2, on a log scale (color). Also shown are the August–October (ASO) mean mixed-layer depth (MLD) for 1999 (pink line) and 1995–2011 mean and variations by one standard deviation (pink dashed and dash-dotted lines) and the annual mean MLD for 1999 (green line) and 1995–2011 mean and standard deviations (green dashed and dash-dotted lines).

45°S close to South America [49.5°W to 75°W]), as we wished to concentrate on the main mode water pools. This mask as calculated for 1999 is shown by the black dotted line in Figure 3a. The objective function regions, referred to throughout as MWFRs, show a clear seasonal cycle in heat content (Figure A1).
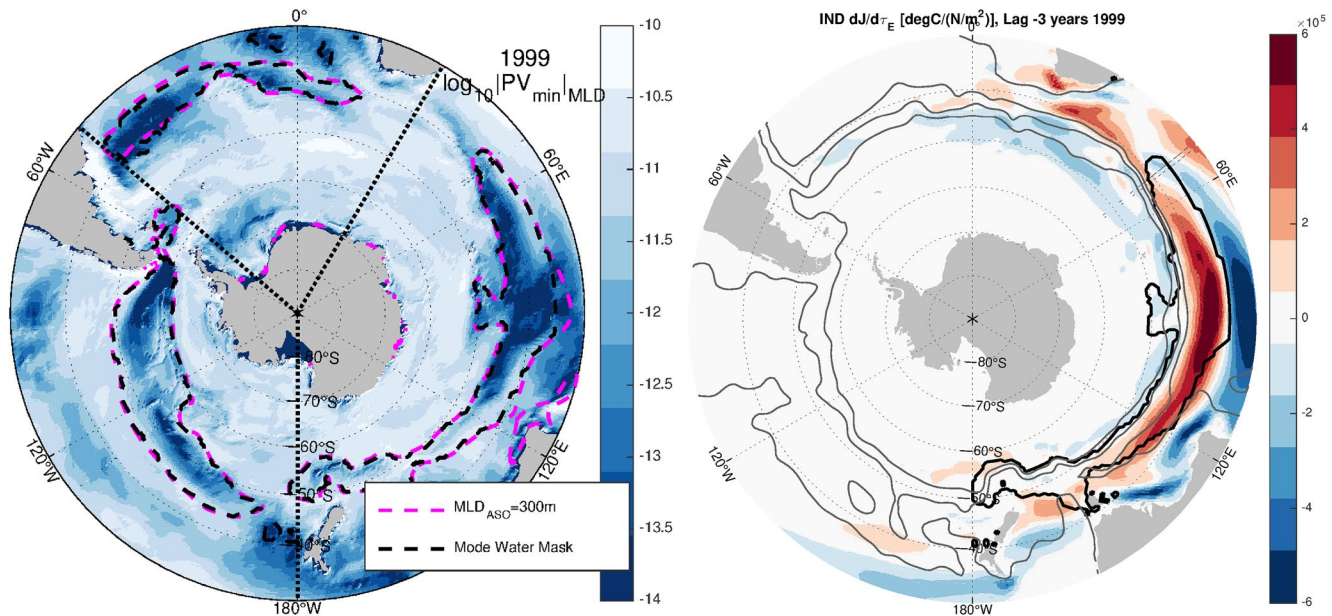

Figure 3. (a) The winter mixed layer encloses mode water formation pools laterally: Blue colors are the absolute value (on a $\log_{10}$ scale) of the 1999 minimum PV at the annual mean mixed-layer depth (the green dash-dotted line in Figure 2). Also shown are the 300 m August–October (ASO) mean mixed-layer depth contour (pink dotted line) and the extent of the mode water mask (black dashed line), as described further in the text. The domain is also divided into three basins by the three longitudinal black dotted lines shown, into the Atlantic, Indian, and Pacific basins referenced throughout. (b) An example sensitivity field: Colors indicate the adjoint sensitivity of the 1999 Indian mode water formation region (MWFR) heat content to zonal wind stress at ~3 years lag. The gray contours indicate the −17, 0, and 30 Sv mean barotropic streamlines, for the entirety of ECCOv4 r2, chosen to highlight the boundary between the ACC and the sub-tropical gyre structure.

We split the SO into three basins using the three latitudinal black dashed lines shown in Figure 3a, and calculate a separate objective function for each basin:

$$J_{SO}^Y = J_{Atl}^Y + J_{Pac}^Y + J_{Ind}^Y, \tag{1}$$

where J_b^Y is the objective function in the given basin b in year Y. The Indian and Pacific basins are divided at 180°W, the Pacific and Atlantic at 49.5°W and the Atlantic and Indian at 30.5°E. Because the adjoint model calculates linear sensitivities, the total SO sensitivity to a given model variable will be the sum of the sensitivities for each basin, that is,

$$\frac{\partial J_{SO}^Y}{\partial X}(\underline{r},t) = \frac{\partial J_{Atl}^Y}{\partial X} + \frac{\partial J_{Pac}^Y}{\partial X} + \frac{\partial J_{Ind}^Y}{\partial X}, \tag{2}$$

where $\partial J_b^Y / \partial X(\underline{r},t)$ is the linear adjoint sensitivity of the objective function J_b^Y to model variable X at point $\underline{r} = (x,y,z)$ and time t.

2.2.2. Objective Function Definition

We re-calculated the objective function based on the same MLD and minimum PV criteria for each of the 20 years in ECCOv4 r2. We chose the annual maximum winter mixed layer depth as the vertical extent of our objective function [denoted $\max(MLD_{ASO})$]. To capture the peak of mode water formation, we chose our objective function to extend to the two months on either side of the peak heat contents of the three basin volumes, that is, from July to November (see Figure A1). Thus, our full objective function for a given year and basin is defined as the following volume averaged heat content:

$$J_b^Y = \frac{1}{V_b^Y \Delta t} \int_{Jul}^{Nov} \iint^{f_b(x,y)} \int_{z=0}^{\max(MLD_{ASO})} \rho_0 c_p \theta(\underline{r},t) \, dt \, dx \, dy \, dz, \tag{3}$$

where $V_b^Y = \iint^{f_b(x,y)} \int_0^{\max(MLD_{ASO})} dx \, dy \, dz$ is the control volume in year Y and basin b, Δt is the averaging time interval, $f_b(x,y)$ is the horizontal mask in basin b; ρ_0, a reference density; c_p, the heat capacity of sea water; and θ, the potential temperature. Note that the extent of the objective function region is calculated offline and so is a fixed volume. The effect of choosing our objective function as defined above, with the lateral extent limited using our mask, rather than just looking at the entire SO mixed layer, is briefly investigated in Section A2.

An example sensitivity field, the sensitivity of the 1999 Indian MWFR heat content to zonal wind stress at ~3 years lag, can be seen in Figure 3b. Thus, red (blue) colors indicate where an increase (decrease) in zonal wind stress in 1996 would result in an increase in the Indian MWFR heat content in 1999. The sensitivity has been scaled by $1/\rho_0 c_p$, and thus units indicate the number of degrees C the similarly scaled MWFR heat content would rise if the zonal wind stress changed by 1 N/m^2.

To assess inter-annual variability in objective function sensitivity, we carried out an ensemble of 13 eight-year adjoint runs, with objective functions defined in each winter from 1999 to 2011. For each ensemble member, sensitivities were output at two week intervals as averages over those two weeks. The sensitivities shown in Figure 4 are ensemble averages of winter (July–September) averages, which are then multiplied by a representative scalar standard deviation for the surface property σ_0 (these values can be found in Table 1) and scaled by $1/\rho_0 c_p$. This makes the units of sensitivity the amount by which a unit perturbation of the given surface property at the relevant point in space and time would raise the objective function J_b^Y in °C.

3. Adjoint Sensitivities to Surface Properties

Winter sensitivity maps (Figure 4) highlight the time when sensitivities peak (Figure 5). Standard deviations calculated over the ensemble sensitivities show that ensemble member variation is largely within the magnitude of the sensitivity and not the location of the sensitivity, that is, there is variation in the magnitude of the sensitivity but not its structure (figures S3–S6). Areas with close to zero mean sensitivity also
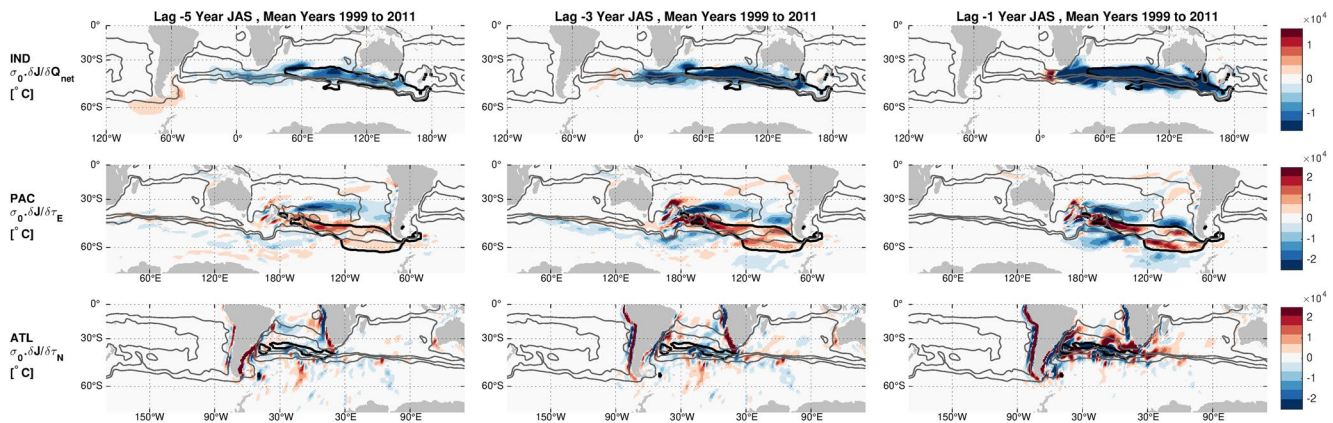
Figure 4. Example sensitivity fields showing the range and general properties of adjoint model simulations: Ensemble mean winter (July–September) sensitivities for surface properties at lags of 5, 3, and 1 years (left, middle, and right columns respectively). The upper row shows sensitivities of the Indian mode water formation region (MWFR) (median location indicated by black contour) to surface net heat flux Q_{net}. The middle row shows sensitivities of the Pacific MWFR (median location indicated by black contour) to zonal wind stress τ_E. The lower row shows sensitivities of the Atlantic MWFR (median location indicated by black contour) to meridional wind stress τ_N. The gray contours indicate the -17, 0, and 30 Sv mean barotropic streamlines. The associated ensemble standard deviations and sensitivities for the basins not shown here can be found in the supporting information.

have close to zero standard deviation. Ensemble averaging therefore highlights the consistent structures in the sensitivity fields, and the year-to-year variability in magnitudes is investigated in Section 3.4. We do not show the fresh water flux sensitivities as they are an order of magnitude smaller than those shown here.

3.1. Sensitivities to Net Heat Flux

Q_{net} is defined as the net heat flux from the ocean to the atmosphere. Thus *negative* sensitivities indicate that a *reduction* in Q_{net}, that is, *less* heat from ocean to atmosphere, results in an increase in the objective function, that is, MWFR heat content, and *positive* sensitivities indicate instead that an *increase* in Q_{net} will result in an increase in the objective function. The largely negative sign of the Q_{net} sensitivities (Figure 4, upper row) is thus not unexpected, showing that a cooling of the ocean surface in these regions results in a cooling of the MWFR. The location of the peak sensitivity is largely on top of, or at previous lags "upstream" of the location of the median objective function, inferred by the expansion of the sensitivities along ACC pathways with increased lag. Again, this is not unexpected and indicates that simply heating/cooling the source waters for the MWFR results in heating/cooling of the MWFR itself. These features are common across sensitivities to Q_{net} for all lags and in each of the three basins (the Pacific and Atlantic can be seen in Figure S4 in the supporting information), and can be used to identify possible source regions of the MWFR waters.

3.2. Sensitivities to Wind Stress

The wind stress sensitivities (Figure 4, middle and lower rows) have a very different structure to the Q_{net} sensitivities, notably there are significant sensitivities of both signs. Dipole-type structures are common across all such wind stress sensitivities (not just those shown here), with features centered on the boundaries of the objective functions and over source water regions upstream. These types of features we associate with convergence/divergence and thus vertical Ekman pumping/suction of water (Czeschel et al., 2012; Jones et al., 2018; Loose et al., 2020).

Additionally, the sensitivities to zonal wind stress stretch both north into the sub-tropical gyres and south across the ACC for all basins. This indicates a direct connection with the strength of the wind-driven sub-tropical gyres and possible links with ACC transport—an increase/decrease in zonal wind stress could imply an increase/decrease in meridional Ekman

Table 1
Representative Standard Deviations σ_0 Used Throughout, Calculated From the Southern Ocean Mean (S of 30°S) of the ECCOv4 Fields' Standard Deviations

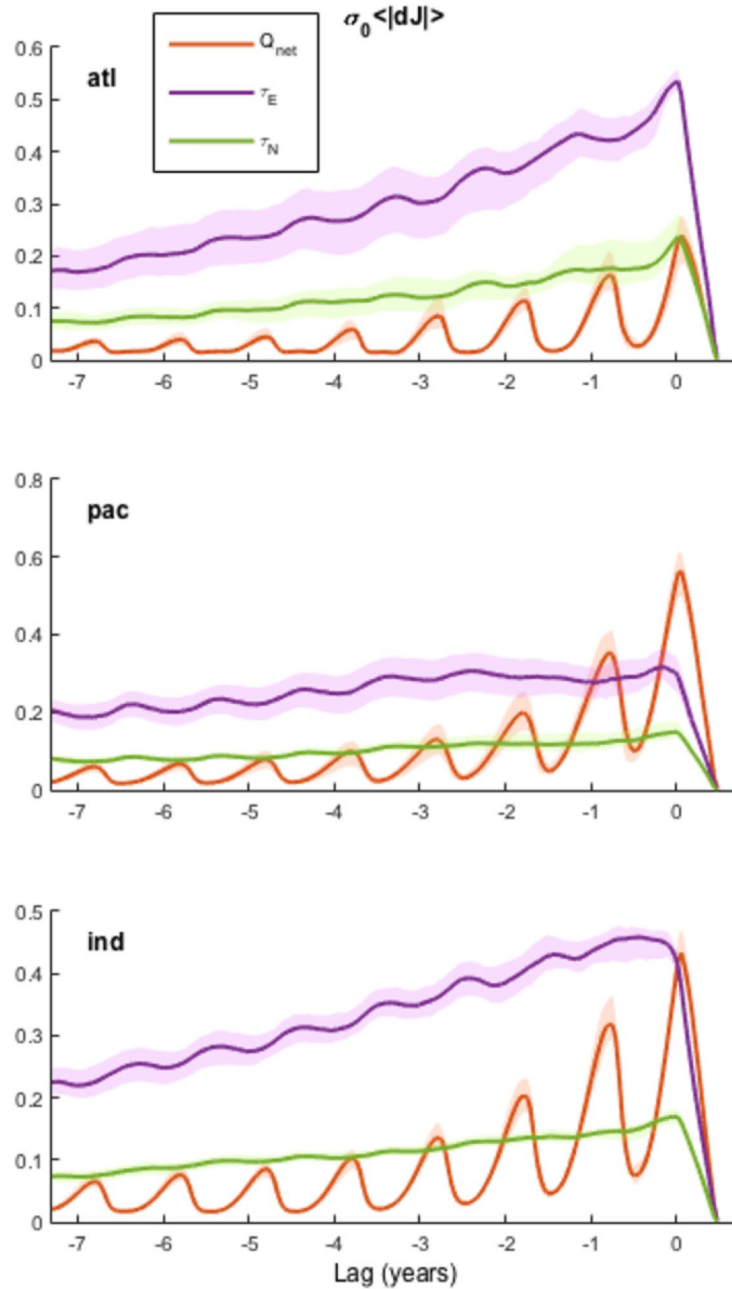| Property symbol	Property name	Standard deviation
Q_{net}	Surface net heat flux	60 W m^{-2}
τ_E	Zonal wind stress	0.08 N m^{-2}
τ_N	Meridional wind stress	0.06 N m^{-2}
Θ	Potential temperature	0.3 °C

Figure 5. Wind stress largely dominates basin-integrated absolute sensitivities: Integrated absolute sensitivities to surface forcings by objective function basin (top to bottom, as labeled), scaled by a representative standard deviation σ_0 and normalized, dimensionless, plotted against lag relative to the start of the objective function. Colors indicate surface net heat flux (Q_{net}, red), and zonal/meridional wind stress ($\tau_{E/N}$, purple/green). The shaded area indicates the ensemble envelope (spanning the ensemble max and min values, *not* a standard deviation or similar) and thick lines the ensemble mean.

transport across the ACC, or a change in the tilt of the isopycnals resulting in a change in zonal ACC transport. Other common features are what appear to be dynamical links with boundary current regions—dynamic because the sensitivities are not in source regions and because the sensitivities often propagate through space over time either along or away from boundaries in patterns similar to topographic, Kelvin, and Rossby waves. This can be seen more easily in the animations provided in the supporting information and is discussed further in Section 4.2.

The negative sensitivity of the Pacific MWFR to zonal wind stress on 1–3 years lags in the region of 120°W to 90°W and South of 60°S (the Amundsen Sea, see Figure 4) is consistent with the results of Close et al. (2013), who find a link between an increased Amundsen Sea Low (ASL, resulting in weaker zonal wind stress) and warmer Sub-Antarctic Mode Water (SAMW). However, this sensitivity is relatively weak compared with zonal and meridional (see Figure S10 in the supporting information) wind stress sensitivities over, to the north of, and upstream of the MWFR, whilst Close et al. (2013) believe the ASL is significant in determining SAMW properties. This may be because although the region shows low sensitivity relative to other regions, the actual wind-stress changes in the region are significantly larger than those in other regions, although this does not appear to be the case for climatological anomalies, see Figure S7.

3.3. Time Evolution of Domain-Integrated Sensitivities

To compare sensitivities over time, we first calculated scaled domain-integrated absolute sensitivities over time for each basin's objective function, that is, the absolute value of the sensitivity is taken before integration over the global domain, meaning positive and negative sensitivities *do not* cancel out. Thus, the integrated absolute sensitivity is the maximum possible impact on the objective function if perturbations are applied with the same sign and magnitude as the sensitivities themselves, demonstrating when the model has the most potential to alter the objective function. In each basin, sensitivity to Q_{net} is highest at lag 0 and then decays with a strong seasonal cycle as the lag increases, peaking each winter (Figure 5). Here, lag 0 is defined as the beginning of the objective function integral, that is, at the start of July—see Equation 3—and so non-zero sensitivities are possible at positive lags between July and the end of the integral in November. Sensitivity to wind stress decays more slowly and has a very slight seasonal cycle, relative to the heat flux sensitivity, which it also appears to be out of phase with.

This study focuses on highlighting possible oceanic mechanisms, but if instead we wished to highlight the origins of forced variability, we could convolve the sensitivities with the contemporaneous anomalies of the surface fluxes from the climatological mean. We have included versions of Figures 4 and 5 weighted by such anomalies in the supporting information Figures S7, and S8.

With our chosen scalings, sensitivity to Q_{net} initially dominates in the Pacific basin, with wind stress sensitivity dominating after around 1-year lag. Wind stress sensitivity dominates in the Atlantic basin, and largely dominates in the Indian basin apart from during the objective function integration period (positive lags), where the Q_{net} ensemble mean sensitivity just dominates. However, the sensitivity that dominates at any given time is dependent on the scaling applied. Scaling the sensitivities instead by the climatological anomaly results in a relative increase in the Q_{net} sensitivity, see supporting information Figure S8, although the overall pattern of Q_{net} sensitivities dominating at short lags (0–1 year) and wind stress dominating at longer lags still holds.

These results indicate that the surface heat flux has the largest impact during winter on mode water formed during that same winter, and thereafter seasonally affects subsequent winters, but to a lesser and lesser degree. The large magnitude of the seasonal cycle means that heat fluxes in past winters have a much stronger influence on MWFRs than intervening summers, even years apart. Wind stress, however, can produce a similar or larger impact than heat flux for years to come, with relatively less seasonal variation, perhaps linked to the dynamical, longer-range nature of the connection with the MWFRs. More explicitly, dynamic processes such as changes in the Ekman pumping over source regions; changes in the ACC or other currents' strengths; the generation of Rossby/Kelvin waves, could influence the MWFRs for many years, regardless of the local mixed layer depth in the MWFR itself. These findings are similar to the results of Jones et al. (2019), who find the heat content of water that subducts from the MWFR is strongly controlled by the sub-tropical gyre strength and structure, which is in turn strongly related to wind-stress over the gyre for the previous 3–4 years.

The integrated sensitivities show remarkable similarity between the basins, despite the different locations and relative sizes of the MWFRs in each basin. The Atlantic MWFR is relatively far north, where it is strongly influenced by the wind-driven Atlantic sub-tropical gyre, which may be why wind stress influences are relatively strongest here. The Pacific and Indian MWFRs are both further south within the ACC, and have

relatively lower sensitivity to wind stress compared with heat flux. In the following section, we investigate the influence of the varying MWFR volumes on the magnitude of the sensitivities.

3.4. Analysis of Links Between Sensitivities and Mixed Layer Depths

The time dependence of the sensitivity to heat fluxes suggests a process very much dominated by mixed layer depths—the sensitivity is largest in the winter when mixed layers are deepest, and the relative importance of past years decreases in time as information from previous winters is lost, with sensitivities at two years lag around half of that at zero years. This is consistent with the fields in Figures 4 (upper row) and S4 that show Q_{net} sensitivities confined to the objective function region (where the mixed layers are deepest) and upstream. The slower decay and relatively weaker seasonal cycle in the wind stress sensitivities also point to the influence of remote processes which are not strongly correlated to local mixed layer depths, and have stronger influences at larger lags.

The link between heat flux sensitivities and mixed-layer depths is further explained by correlations between peak sensitivities and objective function volumes. In each adjoint simulation, the peak in basin-mean absolute dJ/dQ_{net} occurs in July in the lag zero year (see Figure 5), that is, at the beginning of the objective function integration time period (see Equation 3). The magnitude of the peak in each ensemble member is strongly anti-correlated with the objective function volume V_b^Y, with R^2 values given in Table 2. Years with relatively low objective function volumes show relatively large peaks in basin-mean absolute Q_{net} sensitivity, and vice-versa. These anti-correlations are significant at the 99% level, with $R^2 = 0.92$–0.96 across the three basins (see Table 2). This implies that interannual variability in peak sensitivities is almost entirely determined by the volume of the objective function, with larger volumes showing a weaker sensitivity to surface heat fluxes, and vice versa. Given that, at their peak, Q_{net} sensitivities are located directly over the objective function regions (see Figures 4 and S4), this is not surprising: for a given perturbation in surface heat flux, the amount by which a given well-mixed pool will warm will be inversely proportional to the volume of that pool.

There are weaker correlations between peak wind stress sensitivities and objective function volumes, as implied by the weaker seasonal cycles in basin-mean sensitivities. This correlation varies between the three basins—the correlations are strong in the Indian basin, weaker in the Pacific, and only statistically significant when involving meridional wind stress in the Atlantic (Table 2). All correlations are strongest in the Indian basin, with $R^2 = 0.86$ for correlations between objective function volume and peak basin-mean absolute zonal wind stress sensitivity, and $R^2 = 0.69$ for meridional wind stress sensitivity. Each of the peak sensitivities to the Indian MWFR heat content are also strongly correlated with each other (not shown), showing that the objective function volume is a strong control on the magnitude of all three absolute sensitivities to the Indian MWFR. This could be because the Indian basin has the largest volume of the three MWFRs, with a peak climatological heat content over twice that of the Atlantic or Pacific, see Figure A1.

These correlations imply that Q_{net} sensitivities are strongly controlled by changes in objective function volume, which is largely controlled by changing mixed layer depths. The controls on the magnitude of the wind stress sensitivities are not as clear, with the objective function volume a strong control on the absolute magnitude in the Indian basin, but weaker or not significantly correlated in the Pacific and Atlantic. This implies that there are other factors, such as the local density structure, that may be playing an important role in setting the magnitude of the wind stress sensitivities in these basins.

4. Sensitivities to Kinematic and Dynamic Potential Temperature Changes

4.1. Defining Kinematic and Dynamic Sensitivities

As in Marotzke et al. (1999) and Jones et al. (2018), we analyzed the sensitivities of the objective function to potential temperature by splitting it into sensitivities due to changes in temperature along isopycnals (referred to as kinematic changes) and changes in temperature that result in density changes (referred to as dynamic changes). As discussed in Marotzke et al. (1999), this is similar to the decomposition of temperature changes over time into "spice" and "heave" components (Bindoff & Mcdougall, 1994).

The dynamic sensitivity can be written:

Table 2
Strength of Anti-correlations Between Peak Basin-Mean Absolute Sensitivities ($\langle|dJ/dX|\rangle$) and Objective Function Volume (V_b^Y).

	Sensitivity		
Basin	dJ/dQ_{net}	$dJ/d\tau_E$	$dJ/d\tau_N$
Atlantic	**0.92 (0.75–0.98)**	0.15 (0.00–0.60)	**0.65 (0.21–0.88)**
Pacific	**0.94 (0.80–0.98)**	**0.30 (0.00–0.71)**	**0.52 (0.08–0.83)**
Indian	**0.96 (0.87–0.99)**	**0.86 (0.58–0.96)**	**0.69 (0.27–0.90)**
Global	**0.94 (0.81–0.98)**	**0.52 (0.08–0.83)**	**0.33 (0.00–0.73)**

Note: Values shown are R^2 (squared Pearson correlation coefficient) with 95% bounds, bold if significant.

$$F_{dyn} = -\frac{\alpha}{\beta}\left(\frac{\partial J}{\partial S}\right)_\theta, \tag{4}$$

where θ is potential temperature, S is salinity, α is the coefficient of thermal expansion and β the coefficient of haline contraction. The kinematic sensitivity can be written:

$$F_{kin} = \left(\frac{\partial J}{\partial \theta}\right)_S + \frac{\alpha}{\beta}\left(\frac{\partial J}{\partial S}\right)_\theta. \tag{5}$$

Thus, we calculated both dynamic and kinematic sensitivities from the sensitivities to potential temperature and salinity [$(\partial J/\partial \theta)_S$ and $(\partial J/\partial S)_\theta$] output directly from the MITgcm adjoint in combination with the factor α/β calculated from the model output potential temperature on the same two week average time-steps using the TEOS-10 Matlab toolbox (McDougall & Barker, 2011). Note that, unlike the sensitivities to surface fields, each dynamic/kinematic sensitivity snapshot is a three-dimensional field that also depends on depth.

4.2. Kinematic and Dynamic Sensitivity Results

We calculated ensemble mean dynamic and kinematic sensitivities for the same experiments as previously discussed, where the objective function is the heat content of MWFRs. The sensitivities were scaled by $1/\rho_0 c_p$ and so are unitless, that is, the amount by which the objective function would increase in °C for a dynamic/kinematic rise in potential temperature of 1°C. The kinematic sensitivities peak at an average depth of 410 m, and the dynamic sensitivities peak at an average depth of 3 km (not shown), indicating the effectiveness of density changes on the ocean floor for altering pressure gradients (ECCOv4 has a mean depth of 3.8 km in the SO). We choose to plot both quantities at 410 m (Figure 6) as both sensitivities peak close to this depth when scaled by the relevant potential temperature anomalies from climatology (not shown).

Dynamic sensitivities at all depths within the upper ~500 m show a similar structure, with the features seen in Figure 6a (from 410 m depth) moving further away/towards the MWFR regions at longer/shorter lags. Positive dynamic sensitivity indicates that decreasing the density (deepening the density surfaces) at this point would result in an increase in the MWFR heat content, and conversely negative dynamic sensitivity indicates increasing the density (raising the density surfaces) would result in an increase in the MWFR heat content. Within the objective function volume (indicated by the black contours) the sensitivity is largely positive, implying downwelling will produce an increase in the MWFR heat content.

As can be seen with comparison with Figure 6b, much of the strong dynamic sensitivity is placed along the same location as source waters, indicated by strong kinematic sensitivities, but they also stretch further south across the ACC. In the Indian sector, as in the Pacific sector, there are dynamic sensitivities of both signs, both over source regions and extended around these regions. These can be interpreted as highlighting that changes in the strength and structure of the ACC and sub-tropical gyres can draw more or less heat into the mixed layer, although, as previously discussed, any such link would need to be confirmed in a forward run.

In general, dynamic sensitivities for all three sectors are a mix of positive and negative regions, with strong links to continental boundaries. Viewed as animations, one can see that there are many dynamical features that are generated at continental boundaries and then propagate along or away from these boundaries. This can be seen clearly in the Movies S2, S4, and S6, with the main features being:

1. **Atlantic**: A strong dipole directly over the objective function region pattern (as seen in Figure 6a), which rotates in place over time in an anti-clockwise or cyclonic direction, consistent with the westward motion of sensitivity peaks centered at ~30°S and the eastward motion of sensitivity peaks at ~40°S
2. **Pacific**: A strong dipole that is initially centered to the east of New Zealand for 0–1 year lag, which then moves to be centered on for New Zealand at 1–5 years lag (as seen in Figure 6a). A patch of negative
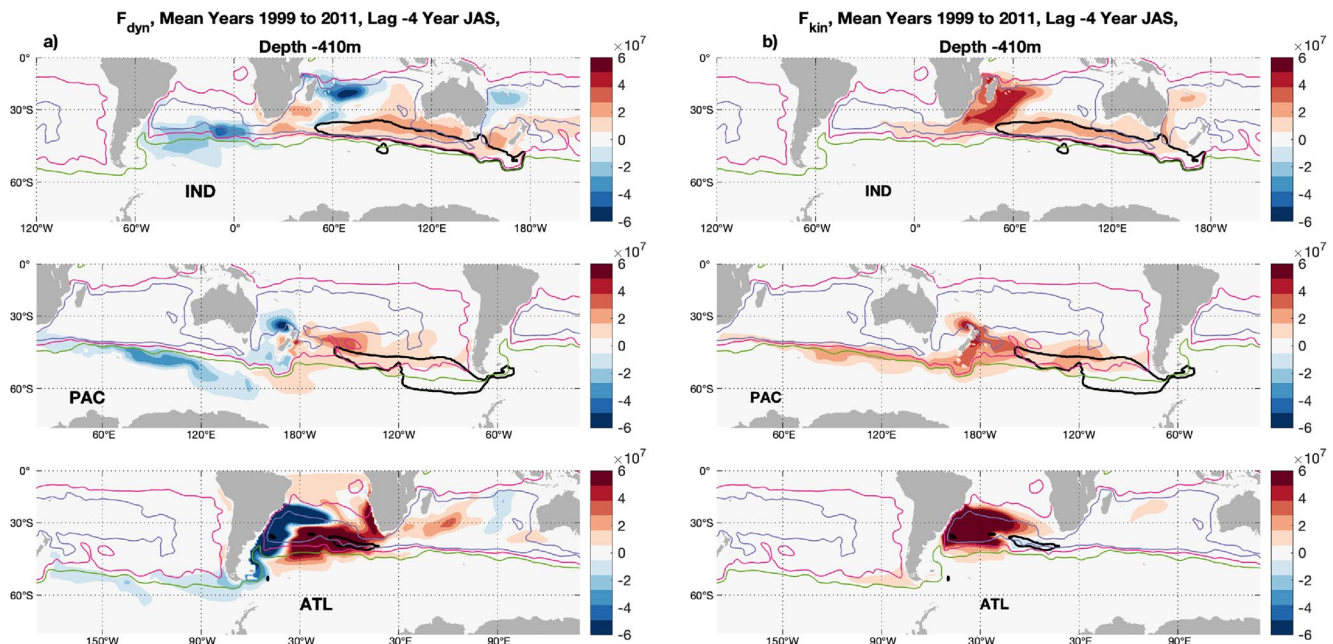
Figure 6. Example dynamic and kinematic sensitivities highlight their different properties: Sensitivities to (a) dynamic and (b) kinematic potential temperature changes at a fixed depth of 410 m, fixed lag of 4 years, in all three basins (top to bottom). The black contour indicates the median location of the objective function at each depth, and as previously, the contours indicate the −17, 0, and 30 Sv mean barotropic streamlines. The associated ensemble standard deviations can be found in the supporting information. Sensitivities are scaled by $1/\rho_0 c_p$ and are unitless.

sensitivity sits upstream along barotropic streamlines. Relatively weaker wave trains are seen to the south of the ACC traveling eastwards, and from the south-west coast of South America traveling westwards

3. **Indian**: For 0–2 years lag, sensitivities are strongest in positive patches along the north of the objective function boundary, in negative patches along the east of South Africa and Australia, and in a wave train traveling eastwards that propagates from below South Africa and then continues just south of the objective function's southern boundary. At longer lags, this wave train can be seen to originate from the eastern boundary of South America, and other westward traveling wave trains can be seen in the Indian and Eastern Pacific oceans at ∼20°S to 30°S

The mean kinematic sensitivities at 4 years lag and 410 m depth, by contrast, are largely single signed (Figure 6b), and sensitivities at shallower depths and at longer/shorter lags are very similar but extend further/less far upstream (see Movies S1, S3, and S5). The Indian and Pacific pools, being close to the northern ACC boundary, are affected by kinematic temperature changes upstream in the ACC, stretching around half its path at 4 years lag. The Indian MWFR is most strongly linked with the Agulhas and Agulhas Return Current regions, as well as more weakly with the East Australian Current region. The Pacific MWFR also shows the strongest links with New Zealand boundary current region. Conversely, the Atlantic pool is shallower (the maximum depth of the median Atlantic MWFR is 480 m, compared with 810 and 910 m for the Pacific and Indian MWFRs, respectively) and further north, more firmly in the sub-tropical gyre, and as such is highly sensitive to local gyre kinematic temperature changes rather than changes in the ACC.

As kinematic temperature changes take place on isopycnals, the sensitivities strongly resemble a passive tracer sensitivity and so reflect the influences of direct heat fluxes or irreversible mixing. In fact, one can directly calculate passive tracer sensitivities in the adjoint model, and they are highly correlated with the kinematic sensitivities at the depths of the objective function (see Figure S10 in supporting information). Thus, kinematic sensitivities reveal approximate source water pathways, and as we consider longer timescales, kinematic sensitivities weaken and are found further away.

4.3. Time Evolution of Domain-Integrated Kinematic and Dynamic Sensitivities

Similarly to Section 3.3, we calculated the domain-integrated absolute dynamic ($\langle |Dy| \rangle$) and kinematic ($\langle |Ki| \rangle$) sensitivities for each basin. We split the integrals into the upper 1000 and 1000 m+ (depths below 1000 m). The summed values are scaled by the maximum J_b^Y for the basin, as well as the total thickness of the integral, to allow for comparison. As with Figure 5, this demonstrates when and where potential temperature changes are most likely to result in changes in the objective function. All three basins show very similar structures, see Figure 7, with the differences being mainly in the timing of the peaks of the various integrals and the degree of variability between ensemble members.

There is relatively more inter-ensemble variability in the Atlantic sensitivities than for the other basins, with several ensemble members showing peaks in upper 1000 m $\langle |Dy| \rangle$ at ~1 year's lag, as seen by the shaded ensemble envelope, whereas the ensemble mean peaks at ~3 years's lag (Figure 7). This increased variability implies a relatively greater state dependence in the Atlantic than the other basins.

The summed absolute dynamic sensitivities are generally an order of magnitude higher than the summed absolute kinematic sensitivities, largely due to the dynamic sensitivities spreading further in space (see Figure 6a and Movies S2, S4, and S6). The magnitude of the thickness-scaled $\langle |Dy| \rangle$ is similar in the upper and lower depth ranges, indicating dynamic pathways within the regions of strongest currents (in the upper 1000 m), are as strong as those at the depths of bottom topography (at 1000 m+). These topographic-depth dynamic pathways with the Pacific and Indian MWFRs are still growing in magnitude at 8 years lag.

The upper 1000 m $\langle |Ki| \rangle$ dominate over the 1000 m+ at all lags, peaking at 0 years and decaying with increased lag, with a slight seasonal cycle apparent. The faster decay in the first few years coincides with the peak sensitivities moving out of the MWFRs and upstream (see Figure 6 and Movies S1, S3, and S5), consistent with passive-tracer-like behavior (see Figure S10).

5. Perturbation Experiments

As discussed in Section 1, we consider adjoint sensitivities to be a useful tool for discovering which regions and timescales are of interest, which can then be explored using fully non-linear perturbation experiments. In this section, we describe how we used the adjoint sensitivities from Section 3 in order to choose the locations for a series of surface forcing perturbation experiments. These perturbation experiments allowed us to directly investigate the full response of our objective functions, including assessing the degree of linearity in the responses, that is, the impact of both the dynamics not captured in the adjoint model and its inherent degree of inexactness.

In the results below, we followed (Verdy et al., 2014) and used the combination of oppositely signed perturbation experiments to calculate the linear and non-linear responses. This allowed for qualitative and quantitative analysis of the two different types of effect, and allowed us to test our assumption that the non-linear component of our objective function is small compared with the linear. Further details of the derivation of the linear and non-linear responses can be found in A3. We applied perturbations in the surface heat flux and the zonal wind stress fields in regions where sensitivities to at least one field were relatively high.

We calculated the integrated heat content of the objective function regions for all perturbation experiments over the fixed maximum winter MLD, following the definition of the objective function J_b^Y:

$$\text{fix}H_b^Y(\theta, \text{MLD}, t) = \iint^{f_b(x,y)} \int_{z=0}^{\max(\text{MLD}_{\text{ASO}})} \rho_0 c_p \theta(\mathbf{r}, t) dx dy\, dz, \tag{6}$$

and thus, the change in heat content with respect to the control simulation (the standard ECCOv4 r2 solution)

$$\Delta\text{fix}H_b^Y(t) = \text{fix}H_b^Y(\theta' - \theta, \max(\text{MLD}_{\text{ASO}}), t), \tag{7}$$
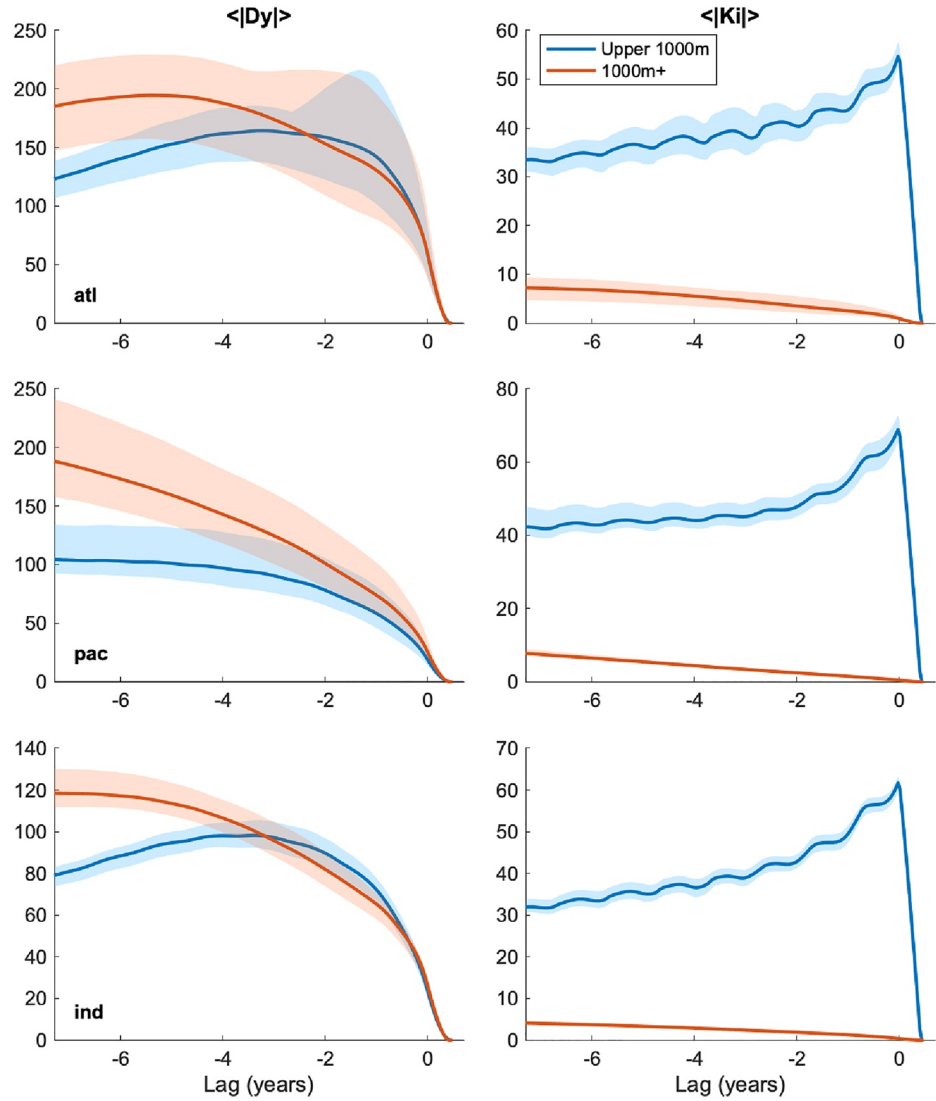
Figure 7. Domain-integrated absolute dynamic sensitivities dominate over domain-integrated absolute kinematic sensitivities, which both grow with time at depth: Domain-integrated absolute dynamic θ sensitivities (left), and domain-integrated kinematic θ sensitivities (right) split by basin (top to bottom, as labeled). Colors indicate the contributions from the upper 1000 m (blue lines), and the depths below 1000 m (red lines). The shaded region indicates the envelope of individual ensembles, and thick lines the ensemble mean. All sensitivities have been scaled by the maximum J_b^Y for that basin and the overall depth of the integral in km, so have a dimension of $\degree C^{-1}\ km^{-1}$.

where θ' is the perturbed simulation potential temperature field and θ is that from the control simulation. The MLD was taken from the control simulation and was therefore the same depth as used in the objective function for the adjoint sensitivity experiments. We also calculated the heat content of the mode water formation regions using the objective function mask for that year, $f_b(x, y)$, but the time-varying *instantaneous* mixed layer depth in each of the simulations:

$$\text{var}H_b^Y(\theta, \text{MLD}, t) = \iint^{f_b(x,y)} \int_{z=0}^{\text{MLD}(t)} \rho_0 c_p \theta(\text{r}, t)dxdydz, \tag{8}$$

and thus the change in the varying-volume heat content

$$\Delta \text{var}H_b^Y(t) = \text{var}H_b^Y(\theta', \text{MLD}', t) - \text{var}H_b^Y(\theta, \text{MLD}, t), \tag{9}$$

where the MLDs were taken instantaneously from the perturbed or control simulations as appropriate. To differentiate between the two volumes, the fixed-volume of the objective function and the instantaneously calculated, varying volume mode water formation region, we refer to them henceforth as the fix-MWFR and var-MWFR, respectively.

5.1. Q_{net} Pacific Perturbation

For our first perturbation experiment, we chose a region in the South-East Pacific identified in other studies as important for downstream SAMW properties (Naveira Garabato et al., 2009), and additionally which shows an interesting pattern of heat flux sensitivity. At two years lag, the Atlantic MWFR heat content has a region of positive sensitivity in this region of the South-East Pacific, just upstream of Drake Passage (see Figure 8a upper panel). This implies that positive heat flux perturbations in this region, that is, increasing heat loss to the atmosphere, will result in a warmer MWFR in the Atlantic in two years time (as previously stated, Q_{net} is defined as positive out of the ocean). Notably there is negative sensitivity over the region of the objective function, so increasing heat loss directly over the Atlantic MWFR would result in a *cooler* MWFR in two years time.

We designed a set of four perturbation experiments to test the sensitivity of the forward nonlinear model to changes in net heat flux in this key region. The black dashed contours in Figure 8a show the region over which the Q_{net} perturbations were applied, in four separate step changes with magnitudes of $\pm 10\,\mathrm{Wm^{-2}}$ and $\pm 100\,\mathrm{Wm^{-2}}$, constant over the box indicated. These perturbations were applied to the forward non-linear ECCOv4 r2 model at the beginning of the model run. Additionally to the changes in Q_{net}, there were resultant changes in the fresh water flux E-P-R, which we do not show because, as demonstrated in Section 3, the sensitivities to this flux are extremely low. Thus the resultant experiment is close to being a test of the influence of Q_{net} independent of other surface fluxes. The perturbation region has a mean Q_{net} of 20 W/m^2 and a seasonal cycle of amplitude 120 W/m^2 in ECCOv4 r2, and so the $\pm 10\,\mathrm{Wm^{-2}}$ perturbations are of similar magnitude to the mean, whereas the $\pm 100\,\mathrm{Wm^{-2}}$ perturbations completely alter the entire seasonal cycle, shifting the region to entirely positive values year-round, or else largely negative.

The perturbation region sits over the Pacific MWFR (see Figure 8a, middle panel), where the sensitivity of the Pacific MWFR heat content is large and negative at all lead times investigated, showing that increasing the heat flux from ocean to atmosphere is an efficient way of cooling this region. At five years lag, the Indian MWFR heat content shows weak positive sensitivity to Q_{net} in the perturbation region (Figure 8a, lower panel). Thus, for a positively signed Q_{net} perturbation in the region indicated, we expect the Atlantic objective function to show an increase in heat content after roughly two years, we expect a decrease in heat content in the Pacific objective function within the first year, and after roughly five years we expect an increase in heat content in the Indian objective function. We expect all these changes to scale linearly with forcing magnitude. The exact adjoint predictions for each year up to 1999 can be calculated by convolving the ensemble mean net heat flux sensitivities with the perturbation, then integrating over time:

$$J_b^{Y'} = \int_Y^{2000} \iint \Delta X . \frac{\partial J_b^Y}{\partial X}\, dxdy\, dt, \tag{10}$$

where $J_b^{Y'}$ is the adjoint prediction of J_b^Y, ΔX is the applied time-constant perturbation in the surface forcing field X, and $\partial J_b^Y / \partial X$ can either be the individual ensemble member sensitivity from a given year Y, or the ensemble mean sensitivity at a given lag (in which case the time integral limits become relative to the beginning of each member simulation, rather than a specific year). The ensemble mean predictions for each of the first eight years (the length of our ensembles) and their standard deviations for each basin can be seen in the thick light blue solid and dashed lines in Figure 8, where the lines span July–November, the objective function period. The prediction for 1999, calculated only using the 1999 ensemble member, is shown similarly in green.

We combined the results of the positively and negatively signed experiments to produce the linear and non-linear impacts for the $\pm 10\,\mathrm{Wm^{-2}}$ and $\pm 100\,\mathrm{Wm^{-2}}$ perturbations. We chose the combinations such that
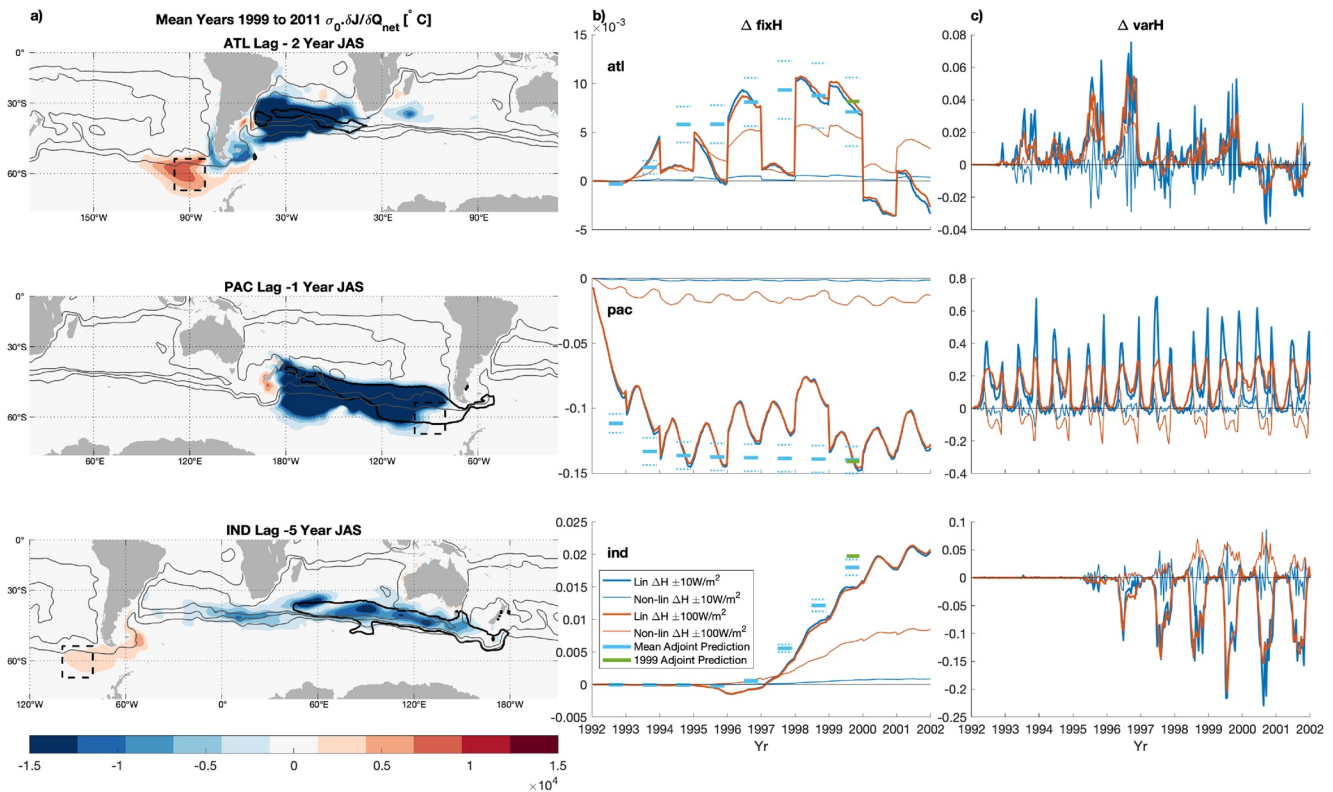
Figure 8. The adjoint sensitivities accurately predict the scaled linear response of the fix-MWFRs heat content: (a) Ensemble mean sensitivities of mode water heat content to Q_{net} in various basins at lags as labeled. Thick gray contours indicated median location of objective functions, black dashed contour indicates location of Q_{net} perturbation (see text for details), gray contours, as before, indicate −17, 0, and 30 Sv mean SSH contours. (b and c) Results of Pacific Q_{net} perturbation experiment, normalized linear (thick lines) and non-linear (thin lines) heat content changes divided by the perturbation magnitude, for either the fix-MWFR (b) or the var-MWFR (c), and for the $\pm 10°$W m^{-2} (dark blue) or $\pm 100°$W m^{-2} (red) experiments. Adjoint predictions for the objective function time periods are shown by thick lines in green (1999 only) and solid and dotted light blue (ensemble mean and standard deviations).

the sign of the linear/non-linear changes indicate the changes for the positively signed Q_{net} perturbations. Note that the heat content changes are discontinuous at the year boundaries due to the changing objective function definition for each year, as the objective function is based on the PV and MLD properties for each individual year, as discussed in Section 2. The magnitude of the changes can be significantly larger for the varying-volume heat contents than the fixed-volumes as the changes in the volume (dependent on the temperature scale used) because changes in the instantaneous MLD result in much larger heat content changes than potential temperature changes alone (see Figures 8b and 8c, noting the different *y*-axis scales.)

One would expect the normalized linear response to be identical for both magnitudes, by definition, and this is largely true, especially for the fixed-volume heat content (see Figure 8b, thick lines, which lie mostly on top of each other). There are small differences at the peaks of the varying-volume responses, likely due to the fact that the bulk formula will have introduced some non-linear changes to the perturbations that will have resulted in the positive- and negative-signed experiments not being exactly symmetric. The non-linear effects (Figures 8b and 8c, thin lines) are smaller in general than the linear effects, but increase in the ± 100 Wm^{-2} case (red lines), as would be expected, becoming almost as large as the linear changes, especially in the Atlantic.

A positive response is seen in the Atlantic (Figures 8b and 8c, upper panels), with both the fix-MWFR and var-MWFR showing linear increases in heat content, starting after roughly two years. The fix-MW-FR response lies within one standard deviation of the ensemble mean prediction for five years out of the first eight (light blue lines), and the exact prediction for 1999 lies very close to the measured response (green line). The heat content of the var-MWFR (Figure 8c) shows large spikes every winter as

the mixed layer deepens, but largely agrees with the sign of the heat content change of the fix-MWFR (Figure 8b).

In the Pacific, at all lags a negative response was expected (Figure 8b middle panel light blue lines), and this is borne out in the fix-MWFR heat content changes (thick red and dark blue lines), which lie within one standard deviation of the ensemble mean prediction for three of the first eight years. However, the sign of the linear change in the var-MWFR (Figure 8c middle panel, bold lines) is opposite to that of the fix-MW-FR: when the heat flux to the atmosphere increases, as in the +10 and +100 Wm^{-2} experiments, the temperature in the fix-MWFR decreases and so does the heat content, but the heat content of the var-MWFR *increases*. This is because the cooler mixed layer deepens, resulting in more net heat content, as can be seen in Figure 9.

The responses in the Indian region (Figure 8b lower panel) are consistent with simple advection downstream—it takes over three years for the effect of the perturbation to reach the Indian region, and it remains much lower magnitude than the Pacific impact. After this, the impact grows year on year, and similarly to the Pacific basin has an opposite-signed linear effect on the fix-MWFR and the var-MWFR. The linear fix-MWFR changes (thick red and dark blue lines) lie within one standard deviation of the ensemble mean predictions (light blue lines) for three of the first eight years. Like the Atlantic, an increase in heat loss to the atmosphere results in an overall warming of the fix-MWFR, and vice-versa. The opposite sign of the response of the fixed and varying volume heat contents is for the same reason as in the Pacific, namely that a warming mixed layer shallows and so decreases its overall heat content when the volume considered is allowed to evolve.

The adjoint predictions lie within one standard deviation from the ensemble mean prediction for just less than half of the winter MWFRs, fewer than would be expected if the ensembles follow a normal distribution. There are a number of possible explanations, including the fact that the years 1992–1998 are not included in our ensemble mean sensitivities, and so can be expected to have slightly different variance. It could also be that the ensemble members do not follow a normal distribution. Whilst the ensemble mean sensitivities did not always predict the correct magnitude, the fix-MWFRs did indeed warm or cool as expected. However, this led to changes in MLD that acted counter to the temperature change and resulted in a larger mixed layer heat content when the mixed layer cooled and a lower mixed layer heat content when the mixed layer warmed (Figure 9). Whilst the temperature change was very linear, the change in MLD had a significant non-linear component, although the linear component is still largest in all but the Atlantic response to the ± 100 Wm^{-2} perturbations (Figure 8c). This is not surprising as the temperature response is strongly linked with the imposed linear Q_{net} changes, whereas the mixed layer response is, as the name suggests, mediated by mixing, which can be non-linear in the case of convective mixing.

5.2. τ_E Pacific Perturbation

We now consider a regional experiment perturbing the zonal wind stress, τ_E. In winter and at three years lag, a clear dipole in the ensemble mean sensitivity of the Pacific MWFR heat content to τ_E can be seen stretching east from New Zealand well into the Pacific (Figure 10a, middle panel). This indicates that a zonal wind stress dipole of this sort, implying downwelling along the dipole center, would produce an increase in the heat content of the objective function region (median location indicated by the solid black contours). A perturbation closely matching this dipole was chosen to test this sensitivity (Figure 10a, dashed black contours) which was applied either imitating the Pacific MWFR heat content sensitivity, with two oppositely signed regions of magnitudes ± 0.1 Nm^{-2}, or with the signs of the two regions reversed. These two perturbations were applied separately as step changes to the forward non-linear ECCOv4 r2 model at the beginning of the model run (the start of 1992). The mean dipole amplitude in ECCOv4 is -0.04 Nm^{-2} with a standard deviation of 0.03 Nm^{-2} in the control run.

Additional to the changes in τ_E, there were resultant changes in the net heat flux Q_{net} due to the bulk formula (not shown). Thus, these experiments are not an exact test of the linear response to the wind-stress perturbations applied, but can nonetheless provide interesting insights into how the linear and non-linear responses compare.
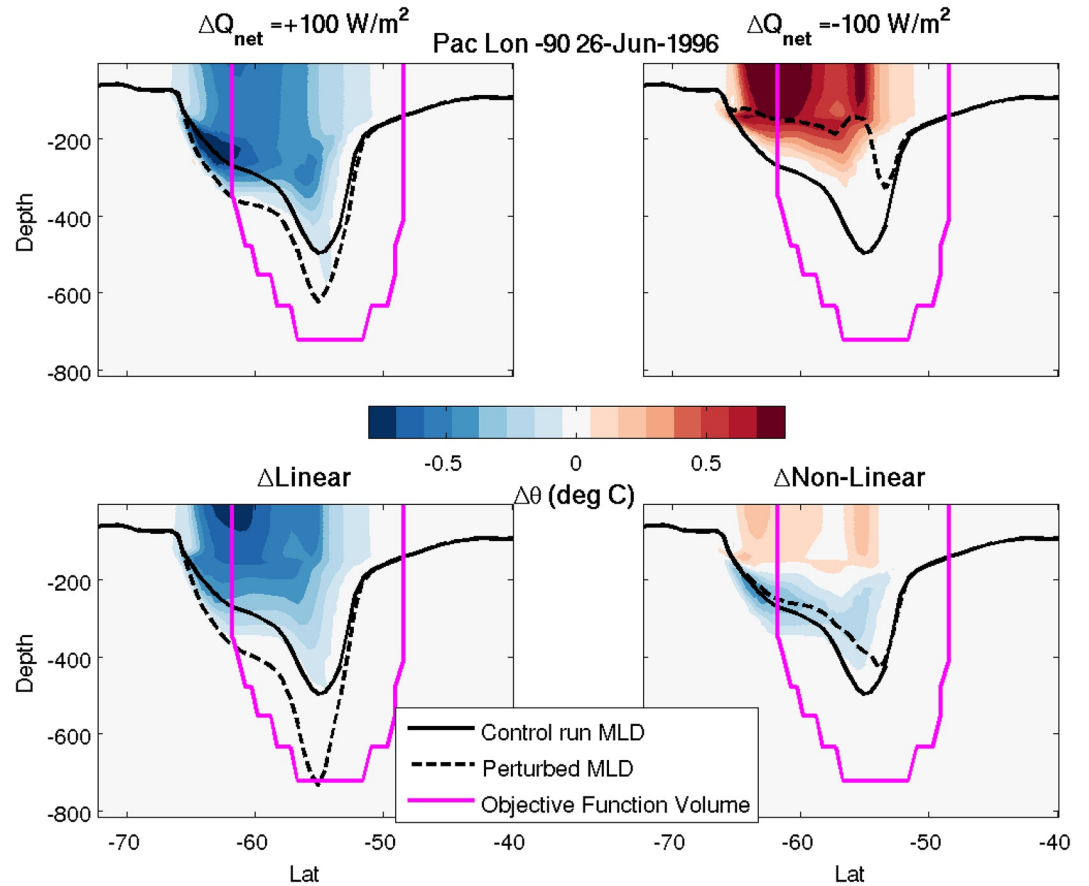
Figure 9. Linear changes in mixed layer depth act counter to linear changes in temperature, leading to opposite changes in heat content of the fix- and var- mode water formation regions (MWFRs): Latitude-depth snapshots of potential temperature changes (color) in the Pacific basin from the Pacific Q_{net} perturbation experiment in June 1996. Q_{net} is, as before, defined as positive from ocean to atmosphere. As labeled, the different panels show the difference from the control run for both positive and negative perturbations, and the combination of these to produce the linear and non-linear changes. The black solid lines show the control run instantaneous mixed layer depth (MLD) and the magenta lines show the 1996 objective function volume (the same in every panel). The black dashed lines show the instantaneous MLD for the perturbation experiments as labeled.

Consistent with the adjoint sensitivity, the linear fix-MWFR heat content in the Pacific sector responded with an increase (decrease) in heat content over time for the positively (negatively) signed perturbation experiment (Figure 10b, thick lines, middle row). The response lies within one standard deviation of the ensemble mean prediction (calculated as in Section 5.1) for six out of the first eight years (very thick and dotted lines). The Indian and Atlantic fix-MWFR heat contents responses are more non-linear than the Pacific, with an especially asymmetric response in the Indian sector, although it becomes more symmetric after 1998.

Note the Atlantic responses are two orders of magnitude lower than climatology (see Figure A1), reflecting its low sensitivity to the perturbation region (Figure 10a, upper row). The Atlantic fix-MWFR responses are of the opposite sign to that predicted by the ensemble mean sensitivities (very thick and dotted lines), demonstrating that the adjoint sensitivities are not appropriate when applying such relatively large perturbations to regions of low sensitivity, when the linear approximation may well be inaccurate.

The Δvar$H$ response (Figure 10b, thin lines), calculated as before from the lateral extent of the objective functions but integrated in depth to the instantaneous MLD, are largely of the same sign as the Δfix$H$ responses in all basins. This is due to large non-linear changes in mixed layer depths in the Pacific and Indian MWFRs (not shown), perhaps related to non-linear Q_{net} forcings via the bulk formula or indicative of
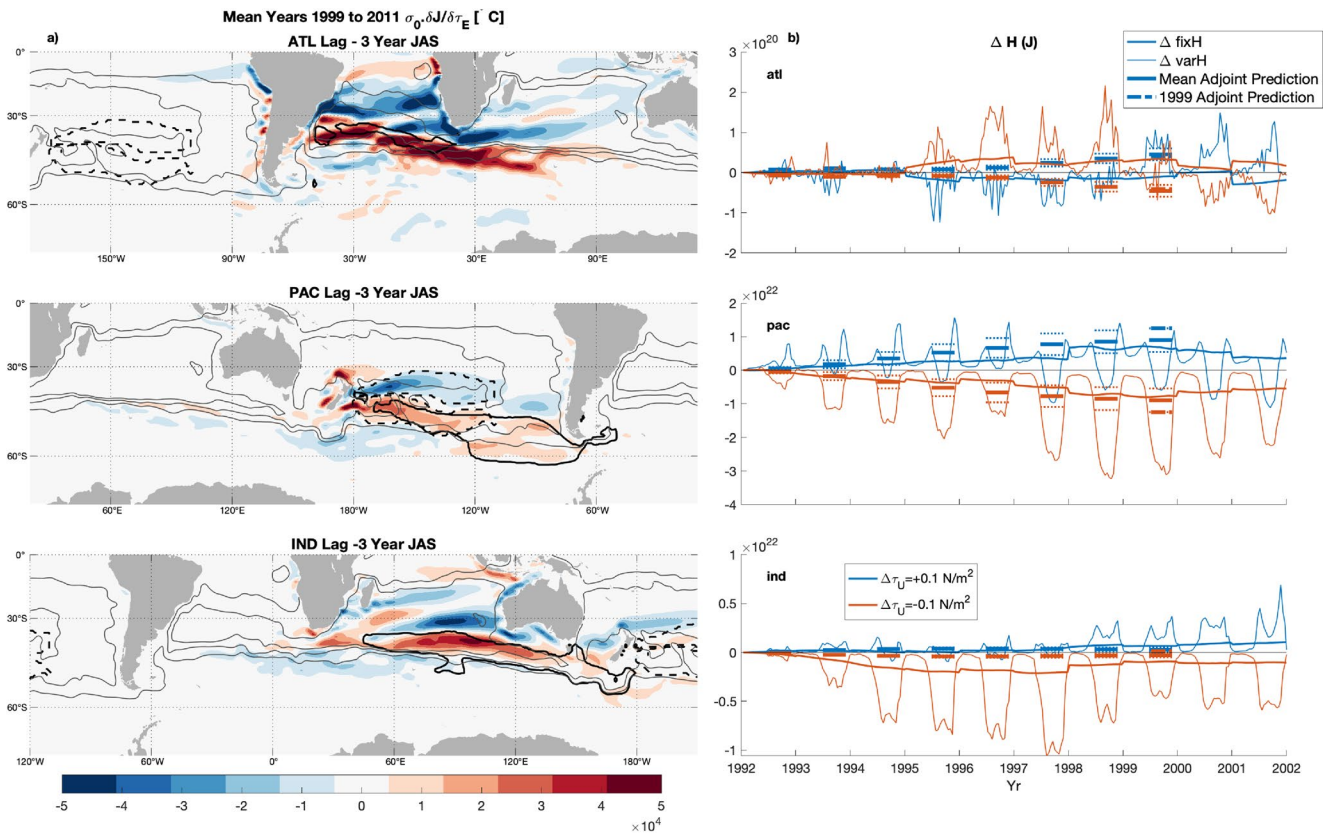
Figure 10. (a) Ensemble mean sensitivities of mode water heat content to τ_E in various basins in winter at 3 years lag as labeled. Black contours indicated median location of objective functions, black dashed contour indicates location of τ_E perturbation (with the positive-signed perturbation matching the sign of the Pacific basin sensitivity shown here), gray contours, as before, indicate -17, 0, and 30 Sv mean SSH contours. (b) Results of Pacific τ_E perturbation experiment. Heat content changes from positively (blue lines) and negatively (red lines)signed perturbation experiments, for either the fix-MWFR (thick lines) or the var-MWFR (thin lines). Adjoint predictions for the objective function time periods are shown by very thick lines either dashed (1999 only), or solid and dotted (ensemble mean and standard deviations).

dynamic processes playing a part in setting the mixed layer depths. In both experiments, there is a seasonal decrease in the Pacific var-MWFR heat content during winter, largely due to non-uniform temperature changes and MLD decreases in the Western lobe of the MWFR.

The results in the Atlantic confirm that perturbing regions with low adjoint sensitivity produces weak linear responses in the forward non-linear model (when compared with regions of significant sensitivity). The relatively poor match to the ensemble mean adjoint predictions is likely due to the inexactness of the adjoint (discussed in Section 6.3), exacerbated by the relatively large perturbation, becoming more apparent when the predicted response is so low, that is, the signal is the same size as the noise.

The results in the Pacific and Indian show that, again, the adjoint sensitivities can accurately predict the linear response of the fix-MWFRs, with a relatively low non-linear response, especially at longer timescales. However, the response of the var-MWFR is highly non-linear, and, in the Pacific, varies spatially within the MWFR. This is consistent with Meijers et al., (2019) who find the East and West Pacific SAMW pools respond differently to forcings.

6. Summary of Results

We have identified locations with properties of winter mode water formation pools within the mixed layer of an observationally constrained model of the SO (Forget, Campin, et al., 2015). Using an adjoint model, we have determined the sensitivity of the fixed-volume heat contents of these mode water formation re-

gions (MWFRs) to surface forcings, changes of potential temperature at constant density, and changes of potential temperature that lead to changes in density, in an ensemble of 13 eight year simulations. These determine the sensitivity of the winter heat content of the MWFRs in the years 1999–2011 to the properties mentioned in previous years. We have highlighted the key aspects of the sensitivities here.

6.1. Summary: Sensitivities to Surface Net Heat Flux and Wind Stress

Analysis of the sensitivity fields revealed that, on the eight year time scale investigated using the ECCO adjoint model, the heat content of the MWFRs is significantly affected by surface net heat fluxes and wind stress, but much less by fresh water fluxes (discussed further on). The heat content of the MWFRs in all three basins was found to be most sensitive to local (within the MWFR), same winter changes to surface heat fluxes, and to both local and remote wind stress changes, which were found to be of comparable integrated magnitude and significant at all lead times.

Heat flux sensitivities have a strong seasonal cycle, with the largest sensitivities occurring during previous winters, with peak values strongly controlled by the objective function volume. This implies that surface heat fluxes are most effective at changing the heat content of MWFRs during winter, when the heat content throughout the deepened mixed layers can be influenced, but that smaller MWFRs allow for greater changes in heat content for a given change in surface heat flux. The mixed layer has a "memory" that allows for changes in one year to affect heat content the next year, indicated by the significant sensitivities in previous winters, although there is a clear decay with time that indicates the influence drops year by year, with the winter peaks in summed absolute sensitivity falling to 10%–15% of the maximum by 8 years lag. The decay of Q_{net} sensitivity with time is likely linked to the location of peak sensitivity moving upstream with increased lag (see Figure 4), where the influence is diluted. This, when combined with local rates of transformation, subduction, and advection results in an overall weakening in integrated sensitivity. These findings extends the role of SAMW formation preconditioning discussed in Sloyan et al. (2010) beyond a single season to over several years. It also aligns well with recent results looking at SAMW variability in the Pacific (Cerovečki et al., 2019; Meijers et al., 2019) who find that while inter-annual variability in SAMW properties is largely the result of local forcing, preconditioning from upstream waters also influences properties on lags of 1–2 years (not unlike in Song et al., 2016).

Wind stress sensitivities revealed dipole patterns, and showed a less pronounced decay in magnitude with time and a less pronounced seasonal dependence, as compared with the heat flux sensitivities. Zonal wind stress sensitivities extend significantly farther south than for other properties, indicating a possible link with ACC dynamics. This is consistent with the findings of Rintoul and England (2002), who find that Ekman transport across the South Antarctic Front (SAF) south of Australia (at roughly 50°S) can drive the variability in T and S properties of SAMW in this region, rather than the variation of surface fluxes. Whilst the volume of the MWFRs shows some influence on the inter-annual variation in peak absolute sensitivities (especially in the Indian basin), the lack of a stronger link is consistent with the fact that wind stresses can influence the heat content of MWFRs through a range of dynamical mechanisms (such as horizontal advection, Ekman pumping/suction, heave of isopycnals) which are not clearly controlled by the volume of the MWFR alone.

6.2. Summary: Sensitivities to Dynamic and Kinematic Potential Temperature Changes

The analysis of sensitivities to surface forcings was supplemented by analysis of the sensitivity of the heat content of MWFRs to potential temperature changes, split into kinematic (at constant density) and dynamic (involving changes in density) components. A summary of the results is provided in Figure 11. Kinematic sensitivities were, for the most part, single-signed and resemble passive tracer sensitivities and thus were largest in direct source regions for the MWFRs, with boundary currents mostly dominating over ACC sources. Dynamic sensitivities showed both signs and indicated the effects of raising/lowering density surfaces.

The largest sensitivities in both cases were over source regions as well as in boundary current regions, across the Southern ACC, and in the sub-tropical gyres. Our results suggest a rich range of possible dynamic pathways can influence the heat content of the MWFRs, which extends widely the regions where accurate
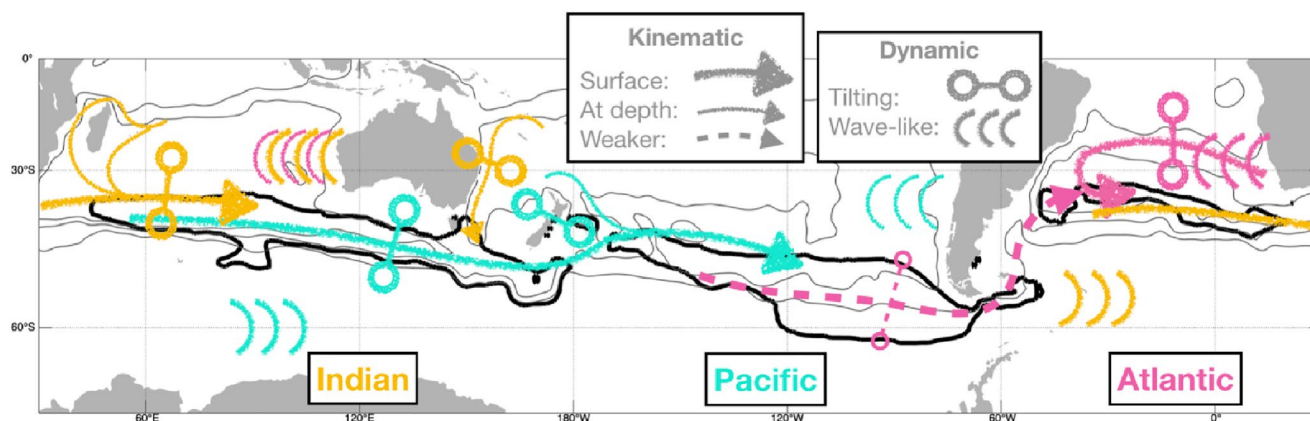
Figure 11. Schematic illustrating the main kinematic and dynamic sensitivities up to ~5 years lag for all three basins: Indian (yellow), Pacific (cyan), and Atlantic (pink). As before, thick black contours show the median location of the mode water formation regions (MWFRs) and gray contours the −17, 0, and 30 Sv mean barotropic streamlines. Arrows indicate paths of kinematic sensitivities, with thinner lines indicating paths only found at depth and dashed lines showing relatively weaker paths. The circles connected by lines indicate where dynamic sensitivities resemble dipoles, where a change in isopycnal gradient will affect the MWFRs (the exact location of the symbols is not meaningful). Groups of curves indicate where wave-like patterns are found.

observations may be required to faithfully model mode water formation regions beyond the local in space and time. When summed over the entire domain and over depth, then scaled by depth range, the dynamic pathways at topographic depths (1000 m+) were of the same magnitude as those at the depths of strongest currents (upper 1000 m). Whilst the sensitivities at the ocean floor are unlikely to be important for observations, due to the relatively weak changes in density at these depths, they may be of relevance for models, showing that small errors in bottom properties could have as much as an impact on the properties of mode water as discrepancies in source regions.

6.3. Summary: Perturbation Experiment Results

Guided by the sensitivity fields, and by previous studies that highlighted regions of relevance for mode water properties, we designed two perturbation experiments using the full forward non-linear model.

These results confirmed that the adjoint sensitivities can successfully predict where and when changes in surface forcings will produce a linear impact on the objective function. In some regions, the sensitivities predicted the overall impact, even for relatively large perturbations, because the non-linear impacts were relatively small. The adjoint sensitivities were accurate at locating regions of high and low linear sensitivity. Additionally, low adjoint sensitivities resulted in low non-linear sensitivities. However, it should be noted that whilst the linearity of the responses were verified, and we compared the responses with the ensemble mean predictions, we did not explicitly verify the exactness of the predictions for each year (see, for example appendix A of Loose et al., 2020). Inexactness can arrive from the approximation to linearity in the adjoint model and the differences between the forward and adjoint models, for example, increasing viscosity in the latter for stability (Forget, Ferreira, & Liang, 2015).

As well as calculating the impact of the perturbations on the fixed-volume MWFRs (fix-MWFRs), we recalculated the volume of the MWFRs in the forward experiments. This allowed us to assess the role played by mixed layer depth variability on the MWFRs through time. These results showed, in some cases, that the varying-volume MWFRs (var-MWFRs) had opposite signed linear heat content changes to the fix-MWFRs. The sometimes significant differences between the fix- and var-MWFRs highlight an important limitation when interpreting the adjoint sensitivities here, computed for fixed volumes, not a water mass or layer which may dynamically alter its thickness in response to forcing.

The zonal wind stress perturbation experiment highlighted the influence of the bulk formula on the surface properties in the model. Whilst linear, opposite-signed perturbations in zonal wind stress were applied in the two experiments, these resulted in significant *non-linear* anomalies in the surface heat flux, due to the

reactions of the bulk formula. In particular, in perturbation experiments of both signs, there was a similar, large decrease in the ocean to atmosphere net heat flux.

6.4. Discussion and Perspectives

It might be surprising to observationalists that there is a lack of strong sensitivities to wind stress or net heat flux south of the ACC (see, e.g., Close et al., 2013). There are a number of reasons why this might be, including that the ECCOv4 model fails to accurately represent the processes responsible for these links in observations, with, for example, too weak off-shelf transport rates (Jones et al., 2019). Additionally, the influence of fresh water fluxes on mode waters has been observed in, for example, Cerovečki et al. (2019); Close et al. (2013), which is not reflected in our results. Salinity changes are likely to have a strong influence on the density and therefore volume of mode waters, but not directly on the temperature of our fixed volume MWFRs.

The adjoint model does not calculate entirely independent sensitivities of the surface forcings considered here (net ocean-to-atmosphere heat flux, wind stress and fresh water flux). The bulk formula couple these quantities together, such that the sensitivities of the net heat flux fields are not entirely independent of wind-driven mechanisms, which can significantly alter the magnitude and spatial patterns of the sensitivity fields (Kostov et al., 2019). The adjoint sensitivities in this study are thus only entirely independent of *non-linear* feedbacks. This makes it harder to compare the results of the perturbation experiments with the adjoint sensitivities, although we expect the non-linear forward model to behave differently than the adjoint linear model. A related issue with ocean-only models is that the bulk formula do not always represent ocean-atmosphere feedbacks correctly (e.g., Strobach et al., 2018), but again the exact magnitude and time scales of this effect is beyond the scope of this work.

An additional way to understand the sensitivities is to convolve them with the contemporaneous anomalies of the surface fluxes from the climatological mean. Rather than showing when the model is most sensitive to changes in in the surface fluxes, this elucidates when and where linear changes to the objective function took place. We have included repeats of Figures 4 and 5 in the supporting information, both of which show similarities with the original figures and do not alter any of our findings.

Combining sensitivity fields with anomaly fields can additionally allow reconstructions of the objective function, in order to attribute the influences of various properties (see, e.g., Pillar et al., 2016). For example, if a particular year had an unusually large MWFR heat content compared with the climatological mean, one could attribute the linear contributions to this difference using the time varying adjoint sensitivities of surface properties convolved with the time varying anomalies of these properties.

The richness of the information contained within the adjoint sensitivities leads to a number of possible uses, many of which we are actively pursuing with collaborators. Most of these possibilities involve combining the adjoint sensitivities with other spatially varying fields. For example, convolving adjoint sensitivities to surface properties with two-dimensional, spatially varying, standard deviation fields highlights where variability is amplified by increased sensitivity. Some regions may instead show high adjoint sensitivity that is offset by low variability. Thus, these analyses highlight where observational campaigns could focus in order to accurately characterize the variability in a given surface forcing. Similarly, predicted changes in surface forcing under climate change scenarios may be expected to have greater impact if they occur over areas of high sensitivity, and the areas of high sensitivity themselves could change as the ocean state changes.

Although care must be taken when interpreting adjoint experiments, specifically considering which timescales and regions that can be expected to have relatively important non-linear effects, the results as presented here indicate the usefulness of adjoint models in producing a rich array of information about regions of interest. Of particular interest to the SO research community are the findings that mode water formation regions appear to be as sensitive to non-local, dynamically linked, wind stress changes on multi-year timescales as to local, kinematically linked, heat flux changes in the same winter. With regards to modeling, it is noteworthy that the adjoint sensitivities can accurately predict the linear behavior of perturbations to the heat content of fixed-volumes in the forward, non-linear model. The exciting range of uses of adjoint sensitivities such as these are just starting to be realized by the community.

Appendix A: Further Details and Derivations

A.1. ECCO Mode Water Formation Region Climatology

Figure A1 shows the climatology of the Mode Water Formation Region heat content from all 20 years of EC-COv4r2 (1992–2011), as defined in Section 2. The Indian MWFR heat content is the largest, peaking at 2.0 $\pm$ 0.2 $\times$ 10^{23} J in September, with the Pacific and Atlantic peaking at 0.9 $\pm$ 0.1 $\times$ 10^{23} J and 0.8 $\pm$ 0.1 $\times$ 10^{23} J, respectively.
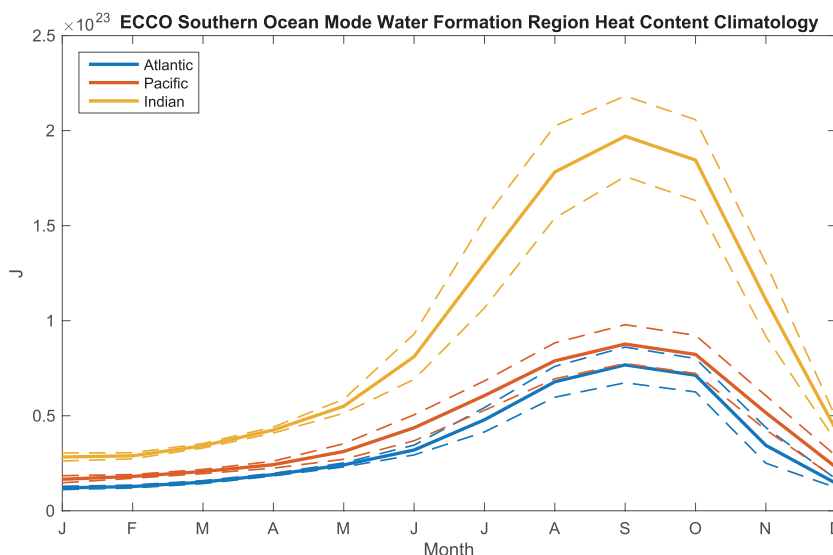

Figure A1. Climatology (1992–2011) of ECCOv4 r2 SO mode water formation region heat content, as defined by our masks in the horizontal, and integrated to the depth of the instantaneous mixed layer, see text for details. Dashed lines show one standard deviation.

A.2. Mask Comparison

Figure A2 shows the domain-integrated absolute sensitivities to surface properties for 1999, comparing the total sensitivity of the 1999 MWFR heat contents as described in Section 2 (red lines) with the sensitivity of the July–November, 1999 maximum mixed layer depth for the whole of the SO (south of 30°S). Thus the difference between the two objective functions is the horizontal extent—the MWFRs are restricted to the areas determined by low PV values and deep mixed layers, whereas the whole SO mixed layer stretches across the domain in the horizontal.

The differences are most striking for the sensitivities to E-P-R, with the mixed layer sensitivities not showing the growth with increased lag that the MWFRs do, however both sensitivities remain extremely small relative to the others calculated. In general, for the net heat flux and wind stress sensitivities, the mixed layer sensitivities peak at a similar or higher value at zero lag, and then decay faster with lag than the MWFR heat content sensitivities. This is not surprising as the SO mixed layer in general has a large surface area and is only on the order of ~100 m depth outside the MWFRs (see, e.g., Figure 2), and so it is expected that it will be most sensitive to recent forcings and quickly lose memory of the past. The absolute wind stress sensitivities in particular show far longer reaching behavior for the MWFRs, likely due to the presence of dipoles along the boundaries of the MWFRs.

This demonstrates that the choice to restrict our objective functions to just the MWFRs themselves produces sensitivities with a richer range of behavior and avoids over-focus on recent surface interactions.
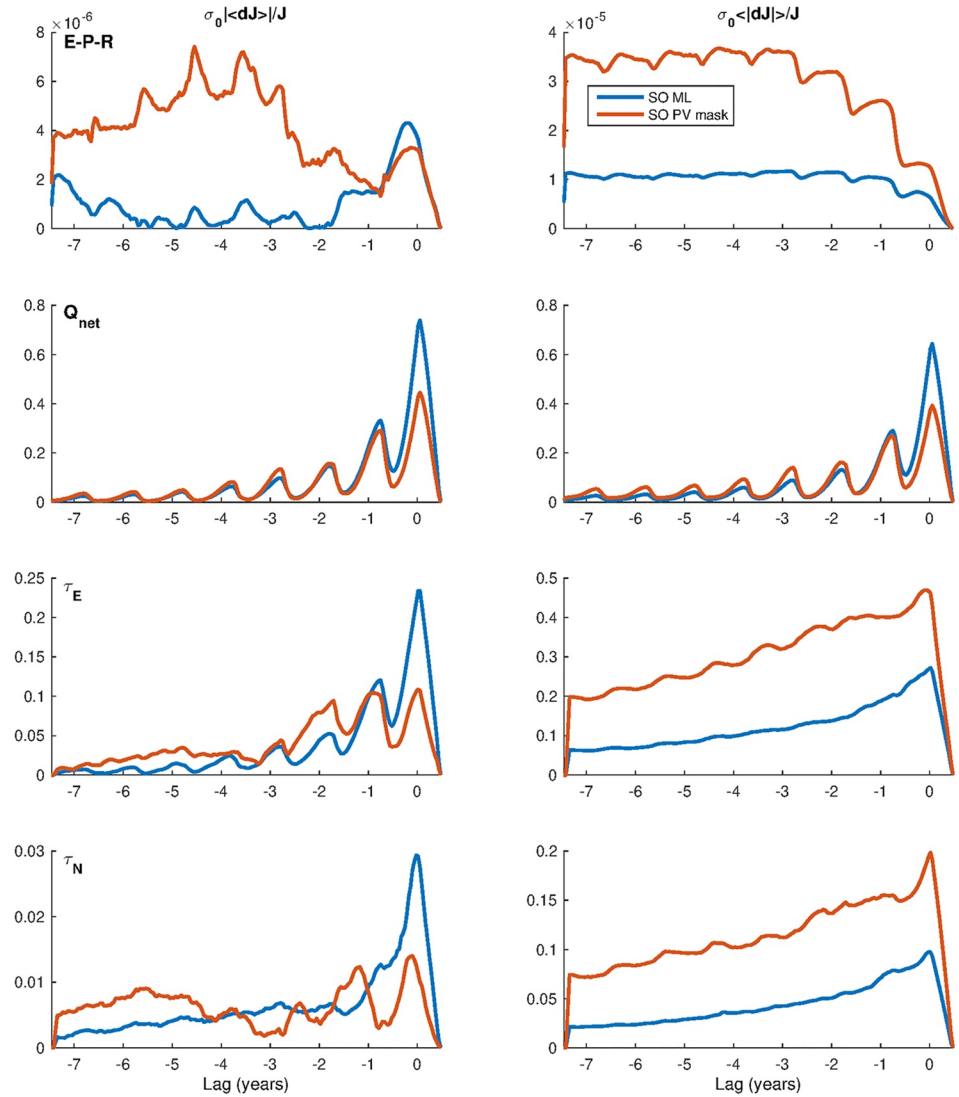
Figure A2. Mean and absolute sensitivities (left and right hand plots respectively) to surface properties as labeled, fresh water flux, net heat flux, zonal, and meridional wind stress, top to bottom. Blue lines show an objective function of the whole SO mixed layer depth July–November, 1999 maximum. Red lines show an objective function of the whole SO 1999 MWFRs—with the horizontal extent determined by the masks described in Section 2 and the vertical extent the July–November maximum mixed layer depth. Sensitivities have been scaled by the representative standard deviations and the value of the objective function J, and then normalized.

A.3. Linear and Non-linear Component Derivation

Given a function $f(x)$ that is infinitely differentiable at a point a, the Taylor series is defined as:

$$f(x) = f(a) + (x - a)\frac{f'(a)}{1!} + (x - a)^2 \frac{f''(a)}{2!} + (x - a)^3 \frac{f'''(a)}{3!} + \dots, \tag{A1}$$

If we assume that a given objective function value J is a function of the model surface forcings, defined by a state vector χ, that is, $J \equiv J(\chi)$, and we consider perturbations to this state vector as $\Delta\chi$, then we can approximate the perturbed objective function as an expansion about the point χ using A1, that is,

$$J(\chi + \Delta\chi) \approx J(\chi) + \Delta\chi J'(\chi) + (\Delta\chi)^2 \frac{J''(\chi)}{2} + \dots, \tag{A2}$$

where we can identify $J'(\chi)$ with the linear component (which is estimated by the adjoint sensitivities $\partial J/\partial\chi$) and $J''(\chi)$ with the non-linear component of $J(\chi)$. Using A1 to similarly define $J(\chi - \Delta\chi)$, we can combine this with A2 to find:

$$\frac{J(\chi + \Delta\chi)}{2} - \frac{J(\chi - \Delta\chi)}{2} \approx \Delta\chi J'(\chi), \tag{A3}$$

$$\frac{J(\chi + \Delta\chi)}{2} + \frac{J(\chi - \Delta\chi)}{2} - J(\chi) \approx (\Delta\chi)^2 \frac{J''(\chi)}{2}, \tag{A4}$$

assuming that $J'''(\chi)$ and higher order terms $\ll J(\chi), J'(\chi)$. Thus, by carrying out the perturbation experiments with state vectors $\chi \pm \Delta\chi$, we can estimate the linear and non-linear behavior of the objective function and test this assumption. We can similarly identify any model variable as a function of the model surface forcings, and use the same method to combine results from the control and perturbation experiments to approximate the linear and non-linear behavior of those model variables.

Data Availability Statement

The ECCOv4-r2 model setup used in this work is available from http://doi.org/10.5281/zenodo.1211363 (Forget, 2018) as an instance of the MIT general circulation model (MITgcm, http://mitgcm.org/). Numerical model runs were carried out on ARCHER, the UK national HPC facility (http://archer.ac.uk/). Adjoint code was generated using the TAF software tool, created and maintained by FastOpt GmbH (http://www.fastopt.com/).

References

Adcroft, A., Campin, J.-M., Hill, C., & Marshall, J. (2004). Implementation of an atmosphere-ocean general circulation model on the expanded spherical cube. *Monthly Weather Review*, *132*(12), 2845–2863. https://doi.org/10.1175/MWR2823.1

Bindoff, N. L., & Mcdougall, T. J. (1994). Diagnosing climate change and ocean ventilation using hydrographic data. *Journal of Physical Oceanography*, *24*(6), 1137–1152. https://doi.org/10.1175/1520-0485(1994)024<1137:dccaov>2.0.co;2

Cerovečki, I., Meijers, A. J., Mazloff, M. R., Gille, S. T., Tamsitt, V. M., & Holland, P. R. (2019). The effects of enhanced sea ice export from the Ross sea on recent cooling and freshening of the Southeast Pacific. *Journal of Climate*, *32*(7), 2013–2035. https://doi.org/10.1175/JCLI-D-18-0205.1

Close, S. E., Naveira Garabato, A. C., McDonagh, E. L., King, B. A., Biuw, M., & Boehme, L. (2013). Control of mode and intermediate water mass properties in Drake Passage by the Amundsen Sea Low. *Journal of Climate*, *26*(14), 5102–5123. https://doi.org/10.1175/jcli-d-12-00346.1

Czeschel, L., Eden, C., & Greatbatch, R. J. (2012). On the driving mechanism of the annual cycle of the Florida current transport. *Journal of Physical Oceanography*, *42*(5), 824–839. https://doi.org/10.1175/jpo-d-11-0109.1

Czeschel, L., Marshall, D. P., & Johnson, H. L. (2010). Oscillatory sensitivity of Atlantic overturning to high-latitude forcing. *Geophysical Research Letters*, *37*(10), L10601. https://doi.org/10.1029/2010gl043177

Forget, G. (2018). *gaelforget/eccov4: Documentation updates*. Zenodo. Retrieved from https://doi.org/10.5281/zenodo.1211363

Forget, G., Campin, J.-M., Heimbach, P., Hill, C. N., Ponte, R. M., & Wunsch, C. (2015). ECCO version 4: An integrated framework for non-linear inverse modeling and global ocean state estimation. *Geoscientific Model Development*, *8*(10), 3071–3104. https://doi.org/10.5194/gmd-8-3071-2015

Forget, G., Ferreira, D., & Liang, X. (2015). On the observability of turbulent transport rates by ARGO: Supporting evidence from an inversion experiment. *Ocean Science*, *11*(5), 839–853. https://doi.org/10.5194/os-11-839-2015

Frölicher, T. L., Sarmiento, J. L., Paynter, D. J., Dunne, J. P., Krasting, J. P., & Winton, M. (2015). Dominance of the Southern Ocean in anthropogenic carbon and heat uptake in CMIP5 models. *Journal of Climate*, *28*(2), 862–886. https://doi.org/10.1175/jcli-d-14-00117.1

Hanawa, K., & Talley, L. D. (2001). Chapter 5.4 mode waters. In Siedler, G., Church, J., & Gould, J. (Eds.), *Ocean circulation and climate* (p. 373–386). Academic Press. Retrieved from http://www.sciencedirect.com/science/article/pii/S0074614201801297

Heimbach, P., Wunsch, C., Ponte, R. M., Forget, G., Hill, C., & Utke, J. (2011). Timescales and regions of the sensitivity of Atlantic meridional volume and heat transport: Toward observing system design. *Deep-Sea Research Part II Topical Studies in Oceanography*, *58*(17–18), 1858–1879. https://doi.org/10.1016/j.dsr2.2010.10.065

Jones, D. C., Boland, E., Meijers, A. J. S., Forget, G., Josey, S. A., Sallee, J. B., & Shuckburgh, E. (2019). Heat distribution in the Southeast Pacific is only weakly sensitive to high-latitude heat flux and wind stress. *Journal of Geophysical Research: Oceans*, *124*(12), 8647–8666. https://doi.org/10.1029/2019jc015460

Jones, D. C., Boland, E., Meijers, A. J. S., Forget, G., Josey, S., Sallée, J.-B., & Shuckburgh, E. (2020). The sensitivity of Southeast Pacific heat distribution to local and remote changes in ocean properties. *Journal of Physical Oceanography*, *50*(3), 773–790. https://doi.org/10.1175/jpo-d-19-0155.1

Jones, D. C., Forget, G., Sinha, B., Josey, S. A., Boland, E. J. D., Meijers, A. J. S., & Shuckburgh, E. (2018). Local and remote influences on the heat content of the Labrador Sea: An adjoint sensitivity study. *Journal of Geophysical Research: Oceans*, *123*(4), 2646–2667. https://doi.org/10.1002/2018jc013774

Khatiwala, S., Tanhua, T., Mikaloff Fletcher, S., Gerber, M., Doney, S. C., Graven, H. D., et al. (2013). Global ocean storage of anthropogenic carbon. *Biogeosciences*, *10*(4), 2169–2191. https://doi.org/10.5194/bg-10-2169-2013

Kostov, Y., Johnson, H. L., & Marshall, D. P. (2019). AMOC sensitivity to surface buoyancy fluxes: The role of air-sea feedback mechanisms. *Climate Dynamics*, *53*(7–8), 4521–4537. https://doi.org/10.1007/s00382-019-04802-4

Landschützer, P., Gruber, N., Haumann, F. A., Rödenbeck, C., Bakker, D. C. E., Van Heuven, S., et al. (2015). The reinvigoration of the Southern Ocean carbon sink. *Science*, *349*(6253), 1221–1224. https://doi.org/10.1126/science.aab2620

Le Quéré, C., Andrew, R. M., Friedlingstein, P., Sitch, S., Pongratz, J., Manning, A. C., et al. (2018). Global carbon budget 2017. *Earth System Science Data*, *10*(1), 405–448. https://doi.org/10.5194/essd-10-405-2018

Levitus, S., Antonov, J. I., Boyer, T. P., Baranova, O. K., Garcia, H. E., Locarnini, R. A., et al. (2012). World ocean heat content and thermosteric sea level change (0–2000 m), 1955–2010. *Geophysical Research Letters*, *39*(10), L10603. https://doi.org/10.1029/2012gl051106

Loose, N., Heimbach, P., Pillar, H., & Nisancioglu, K. (2020). Quantifying dynamical proxy potential through shared adjustment physics in the North Atlantic. *Journal of Geophysical Research: Oceans*, *125*(9), e2020JC016112. https://doi.org/10.1029/2020jc016112

Lumpkin, R., & Speer, K. (2007). Global ocean meridional overturning. *Journal of Physical Oceanography*, *37*(10), 2550–2562. https://doi.org/10.1175/JPO3130.1

Marotzke, J., Giering, R., Zhang, K. Q., Stammer, D., Hill, C., & Lee, T. (1999). Construction of the adjoint MIT ocean general circulation model and application to Atlantic heat transport sensitivity. *Journal of Geophysical Research*, *104*(15), 29529–29547. https://doi.org/10.1029/1999JC900236

Marshall, J., & Speer, K. (2012). Closure of the meridional overturning circulation through Southern Ocean upwelling. *Nature Geoscience*, *5*(3), 171. https://doi.org/10.1038/ngeo1391

Mazloff, M. R., Heimbach, P., & Wunsch, C. (2010). An eddy-permitting Southern Ocean state estimate. *Journal of Physical Oceanography*, *40*(5), 880–899. https://doi.org/10.1175/2009jpo4236.1

McDougall, T. J., & Barker, P. M. (2011). *Getting started with TEOS-10 and the Gibbs Seawater (GSW) oceanographic toolbox* (p. 1–28) SCOR/IAPSO WG127.

Meijers, A. J. S., Cerovečki, I., King, B. A., & Tamsitt, V. (2019). A see-saw in Pacific subantarctic mode water formation driven by atmospheric modes. *Geophysical Research Letters*(*46*), 13152–13160. https://doi.org/10.1029/2019GL085280

Mikaloff Fletcher, S. E., Gruber, N., Jacobson, A. R., Doney, S. C., Dutkiewicz, S., Gerber, M., et al. (2006). Inverse estimates of anthropogenic CO_2 uptake, transport, and storage by the ocean. *Global Biogeochemical Cycles*, *20*(2). https://doi.org/10.1029/2005gb002530

Naveira Garabato, A. C., Jullion, L., Stevens, D. P., Heywood, K. J., & King, B. A. (2009). Variability of subantarctic mode water and Antarctic intermediate water in the Drake Passage during the late-twentieth and early-twenty-first centuries. *Journal of Climate*, *22*(13), 3661–3688. https://doi.org/10.1175/2009JCLI2621.1

Newman, L., Talley, L., Mazloff, M., Galton-Fenzi, B., Ackley, S., Heimbach, P., et al. (2015). *Southern Ocean community comment on the year of polar prediction implementation plan, SOOS report series (Tech. Rep. No. 2)*. Retrieved from http://www.soos.aq/

Pillar, H. R., Heimbach, P., Johnson, H. L., & Marshall, D. P. (2016). Dynamical attribution of recent variability in Atlantic overturning. *Journal of Climate*, *29*(9), 3339–3352. https://doi.org/10.1175/JCLI-D-15-0727.1

Rintoul, S. R., & England, M. H. (2002). Ekman transport dominates local air-sea fluxes in driving variability of subantarctic mode water. *Journal of Physical Oceanography*, *32*(5), 1308–1321. https://doi.org/10.1175/1520-0485(2002)032⟨1308:ETDLAS⟩2.0.CO;2

Roemmich, D., Church, J., Gilson, J., Monselesan, D., Sutton, P., & Wijffels, S. (2015). Unabated planetary warming and its ocean structure since 2006. *Nature Climate Change*, *5*(3), 240. https://doi.org/10.1038/nclimate2513

Sallée, J.-B., Speer, K., Rintoul, S., & Wijffels, S. (2010). Southern Ocean thermocline ventilation. *Journal of Physical Oceanography*, *40*(3), 509–529. https://doi.org/10.1175/2009JPO4291.1

Sloyan, B. M., Talley, L. D., Chereskin, T. K., Fine, R., Holte, J., Sloyan, B. M., ...Holte, J. (2010). Antarctic Intermediate Water and Subantarctic Mode Water formation in the southeast Pacific: The role of turbulent mixing. *Journal of Physical Oceanography*, *40*(7), 1558–1574. https://doi.org/10.1175/2010jpo4114.1

Smith, T., & Heimbach, P. (2019). Atmospheric origins of variability in the South Atlantic meridional overturning circulation. *Journal of Climate*, *32*(5), 1483–1500. https://doi.org/10.1175/jcli-d-18-0311.1

Song, H., Marshall, J., Follows, M. J., Dutkiewicz, S., & Forget, G. (2016). Source waters for the highly productive Patagonian shelf in the southwestern Atlantic. *Journal of Marine Systems*, *158*, 120–128. https://doi.org/10.1016/j.jmarsys.2016.02.009

Strobach, E., Molod, A., Forget, G., Campin, J.-M., Hill, C., Menemenlis, D., & Heimbach, P. (2018). Consequences of different air-sea feedbacks on ocean using MITgcm and MERRA-2 forcing: Implications for coupled data assimilation systems. *Ocean Modelling*, *132*, 91–111. https://doi.org/10.1016/j.ocemod.2018.10.006

Talley, L. D. (1999). Some aspects of ocean heat transport by the shallow, intermediate and deep overturning circulations. *Geophysical Monograph Series*, *112*, 1–22. https://doi.org/10.1029/gm112p0001

Verdy, A., Mazloff, M. R., Cornuelle, B. D., Kim, S. Y., Verdy, A., Mazloff, M. R., et al. (2014). Wind-driven sea level variability on the California coast: An adjoint sensitivity analysis. *Journal of Physical Oceanography*, *44*(1), 297–318. https://doi.org/10.1175/jpo-d-13-018.1