Referee #1

We thank Referee #1 for the balanced and thorough review that touches upon multiple aspects of our study and helps us improve our manuscript.

All line numbers in this response refer to the updated manuscript.

This manuscript aims to explore historical variability of Labrador Sea Water volume and the processes that drive that variability. The authors use the ECCO state estimate and its adjoint, in combination with an objective function to reconstruct and attribute variability. Although in the discussion the authors indicate the study has multiple limitations based around the assumption their approach makes, the authors show that the volume of LSW accumulated in the Labrador Sea exhibits a delayed response to surface wind stress and buoyancy forcing outside the convective interior of the Labrador Sea (e.g. the NAC). They use this response to predict a fraction of the LSW variability a year in advance. They also suggest that all of wind stress, freshwater fluxes, and heat fluxes make contributions of comparable magnitude to LSW volume anomalies, rather than wintertime cooling being the dominant driver.

Given the importance of the Labrador Sea and its deep water formation, looking at drivers of its variability is important. The new results, if robust, will help us understand past changes in the Labrador Sea, as well as its future evolution. The paper is also generally well written, with good quality figures. Thus, the work is eventually worth publishing.

Thank you for pointing out the positives and the strengths of the manuscript!

Before then, I do have some concerns about the authors approaches and assumption that need to be addressed before it can be published. I also think the background literature needs to be improved and the work better sited in what we already know about the Labrador Sea.

We have expanded the literature review in the introduction (lines 32-35, 41-59, 69-77, 124-130, 135-140).

Thus. I would recommend major revisions.

The introduction feels short and underwhelming for such a full length manuscript. There is less than a page of general material before the authors begin to delve into their own approach and plans for the manuscript.

We have expanded the literature review in the introduction (lines 32-35, 41-59, 69-77, 124-130, 135-140). The introduction is now more than twice as long as it was before.

I would like to see more background on the Labrador Sea and LSW formation. What do we know about the processes that drive its formation and variability to help put the author's work in context. Especially important is a well-developed literature over the past ~20 years looking at processes that drive LSW formation, from idealized models, to water mass transformation approaches, to numerical models.

We have expanded the literature review on the topic of LSW formation in the newly added lines 41-59

Many of these studies have looked at the role of winds,

We have discussed the role of winds in lines 55-59 of the introduction.

the buoyancy forcing (including separating the two components – the heat and freshwater forcing) – yet I don't see a single reference to any such studies.

We refer to previous literature on the role of surface cooling and freshwater forcing in the newly added lines 41-52 in the introduction.

And although the authors focus is on air-sea forcing, lateral exchange from the boundary currents to the interior are also important in balancing the air-sea forcing, yet I don't see any discussion of these processes and their potential relevance.

We have added references and discussed lateral exchange in the newly added lines 57-59 and 72-74.

Also, later on the authors bring in the role of forcing in regions like the EGC, the NAC, but yet there really isn't a discussion of the sub-polar gyre circulation and how all these pieces are linked.

Later on, in lines 478-485 we have now emphasized the role of particular currents for transporting temperature, salinity, and density anomalies. We have also pointed more explicitly to our schematic diagram of ocean currents (now separated as a standalone Fig. 1).

To truly understand the significance of the authors work and what is now, a reader needs that missing background.

We have expanded the background review in the introduction (lines 32-35, 41-59, 69-77, 124-130, 135-140).

The authors use the ECCO state estimate for their analysis. I feel the authors assume the reader is very familiar with this product. Which is likely not always the case. More background is needed, especially discussion of the quality of its representation of the SPG and the Labrador Sea, so give the reader confidence in its being the authors underlying tool.

We have expanded our description of ECCO in lines 136-140 and 150-162, including details about the framework (lines 150-158) and the quality of its representation of the SPG and Labrador Sea (lines 158-162).

Also, what years were ECCO run over? I understand it limits how far back the authors can go. But why does their study stop in 2012. The latest version of ECCO goes to 2017 I believe. And given the strong convection in the Labrador Sea (and its shift to the east) in 2012, and 2015-2018, it seems to be a major loss to ignore those recent years.

In the first submission of this manuscript, we used ECCOv4 release 2, which ended in December 2011. Following your comment, we have redone our calculations using the subsequent ECCOv4 release 3 and release 4. In the updated manuscript, we show results from ECCOv4 release 4

which gives us a timeseries of LSW variability from 1992 until the end of 2017 (e.g., Fig. 2, Fig. 3).

Another important question is the authors definition of LSW. How was the range 27.7-27.84 arrived at? It doesn't seem to fit any common definition I've seen in the literature. Many studies break it down into upper LSW and classical LSW, but those ranges are typically 27.68-27.74 and 27.74-27.8. The various Yashayaev papers argue there are various 'vintages' of LSW and one can't use a fixed density range through time (Feucher et al show how much difference that can make in a model that include salinity drift). Other works (especially models) define a higher upper bound because of those drift (i.e. a model LSW). Do the authors pick their LSW range based on ECCO's behavior? If so, that is fine, but the choices must be explained and justified, with discussion of why the range differs from other studies. I'd also like to see some sensitivity analysis related to that range. And then discussion of how such a range may impact the results (in the discussion section).

As we now say in lines 124-127, "Our main constraint for defining LSW is based on vertical stratification (a component of the potential vorticity, PV), while we also define generous potential density bounds to help identify the watermass. This is similar to the definition of LSW used in Li et al. (2019).."

We have cited Feucher et al. in line 185 (and our updated manuscript also cites 5 papers where Yashayaev is a first author and several co-authored papers on LSW).

Most importantly, we have elaborated on the choice of density range in lines 170-192, where we state:

"Following Talley and McCartney (1982), Zou and Lozier (2016) and Li et al. (2019), we approximate PV in terms of the vertical stratification:

$$PV \approx f\frac{N^2}{g} \quad (2)$$

where f is the Coriolis parameter, g is the gravitational acceleration, and N is the buoyancy frequency. Thus, the PV condition which we impose ensures that we define LSW as a weakly stratified watermass. Li et al. (2019) demonstrate that in observations and in most models, this criterion is universal and sufficient for identifying LSW in the Labrador Sea. However, in some models and model configurations, low stratified water can be found below the LSW layer in the basin (Li et al., 2019). Thus, we introduce a second constraint, which sets bounds on the potential density σ_θ of the watermass referenced to the surface:

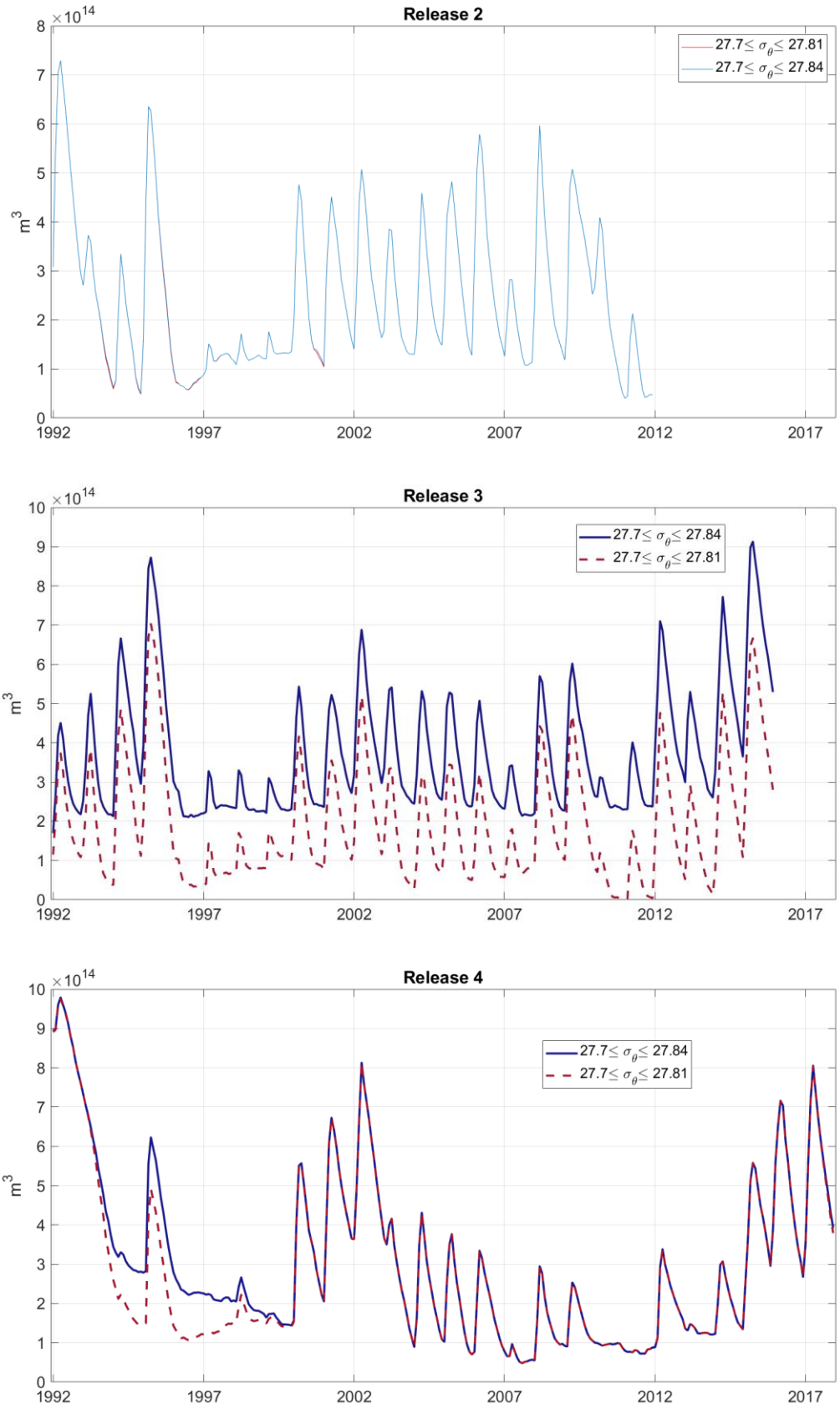
$$\sigma_{\theta\ lower} < \sigma_\theta < \sigma_{\theta\ upper} \quad (3)$$

where $\sigma_{\theta\ lower} = 27.7\mathrm{kg\ m^{-3}}$ and $\sigma_{\theta\ upper} = 27.84\ \mathrm{kg\ m^{-3}}$. This potential density constraint is deliberately very generous because in ECCO, similarly to many models and observations, the density, temperature, and salinity of the LSW formed in the Labrador Sea differ from year to year (Feucher et al., 2019). However, we have tested a stricter density constraint with $\sigma_{\theta\ upper} = 27.81\ \mathrm{kg\ m^{-3}}$. Using $27.81\ \mathrm{kg\ m^{-3}}$ as the upper bound gives the same results for the volume of LSW in release 2 of ECCOv4. In contrast, in the subsequent release 3, the Labrador Sea in ECCO has a secondary deep layer of low-stratified water denser than $27.81\ \mathrm{kg\ m^{-3}}$ and

distinct from the LSW above. The existence of this deep low-stratified water explains most of the time-mean offset between our calculation of LSW volume in ECCOv4 release 3 compared to releases 2 and 4. In the most recent release 4, there is only a brief period between the mid to late 1990s where the model produces and then stores low-stratified water all the way down to $27.84\ \mathrm{kg\ m^{-3}}$ in the Labrador Sea. Outside this period both $\sigma_{\theta\ upper} = 27.84\ \mathrm{kg\ m^{-3}}$ and $\sigma_{\theta\ upper} = 27.81\ \mathrm{kg\ m^{-3}}$ give the same results for the volume of LSW in ECCOv4 release 4."

As described in the new text above, we have done a sensitivity analysis with respect to this density range. This sensitivity analysis would take up too much space if we were to include it in the manuscript or the appendices, but we are providing it here for your reference. The figures below illustrate that in ECCOv4 release 2, $\sigma_{\theta\ upper} = 27.84\ \mathrm{kg\ m^{-3}}$ and $\sigma_{\theta\ upper} = 27.81\ \mathrm{kg\ m^{-3}}$ give almost identical results. In release 4, using $\sigma_{\theta\ upper} = 27.84\ \mathrm{kg\ m^{-3}}$ or $\sigma_{\theta\ upper} = 27.81\ \mathrm{kg\ m^{-3}}$ gives the same result except in the late 1990s. The reason behind this good agreement is that our main criterion for defining LSW is actually PV rather than density. In effect, LSW is defined as water with low vertical stratification. This definition has been promoted by Zou and Lozier (2016) and Li et al. (2019), cited in the manuscript.

In contrast to the other releases, in ECCOv4 release 3, there is a secondary deep layer of low-stratified water, which would correspond to NEADW rather than LSW. This creates a larger discrepancy between using $\sigma_{\theta\ upper} = 27.84\ \mathrm{kg\ m^{-3}}$ or $\sigma_{\theta\ upper} = 27.81\ \mathrm{kg\ m^{-3}}$. This issue, however, is not present in release 2 or release 4.



Additional Figure: Sensitivity analysis of the LSW definition with respect to density, while using the same PV criterion.

I also wonder about the choices of the time period of the winter objective function and the two winter functions. Firstly, today, is it really not possible to broadly compute the functions for all years and seasons. Yes, it would take longer but I would have thought the authors could have gotten some supercomputer access to do so. But if it is not possible to go beyond these 3 periods, I am still wondering about their choice. 2006, 2007 and 2011 are all towards the end of the ECCO timeseries used.

We had limited computational resources early on and currently no access to a computational budget for this project. We have added two more runs with summertime objective functions (e.g., the new Fig.2, Fig. 3, Figure 11), but we cannot extend the ensemble further by considering objective functions in additional years.

In lines 233-237, we have now justified the choice of years towards the end of the ECCOv4 release 2 timeseries for defining objective functions: "The selected three years are close to the end of the ECCOv4 release 2 state estimate because that allows us to compute lagged sensitivity over longer periods leading up to these three years. We use the sensitivity patterns for the ECCOv4 release 2 objective functions to reconstruct not only LSW variability in release 2, but also to attempt reconstructing LSW in the more recent ECCO release 3 (not shown) and in release 4, that extends further in time until the end of 2017."

And although ECCO has significant LSW formation in 2006, I don't believe that year is shown to have high LSW in the observations, such as some of the Yashayaev papers. March 2008 would have been a better strong LSW formation winter.

Unfortunately, March 2008 would not have been a very representative month in ECCOv4. Lines 373-375 now say: "However, all releases of ECCOv4 seem to underestimate the magnitude of the 2008 relative increase in LSW, where the underestimation is most pronounced in release 4."

My next major questions resolves around the forcing functions that the authors look at – wind, heat and freshwater. It looks like the authors are comparing each term to the LSW formation. But for example, sensible heat fluxes depend significantly on the wind. How is that taken into account and attributed? At the very least, this needs to be discussed to help the reader understand what the authors mean by wind or heat forcing, for example.

Following your recommendation, we have first clarified this in lines 119-123:

"We treat the surface boundary conditions in such a way as to avoid overlap and double-counting between their interrelated contributions. For instance, when we analyze the impact of surface winds, we account for their input of momentum. In contrast, the winds' impact on air-sea heat exchange is considered to be part of the heat flux contribution to LSW variability. Similarly, when we talk about the effect of surface heat fluxes, we do not include their impact on surface salinity via evaporation, as that is accounted for in the contribution of surface freshwater fluxes."

We have also expanded the paragraphs in lines 272-279 that discusses this. We now clarify

"Therefore, both the net surface heat flux Q and the default sensitivity patterns $S(Q, x, t - t_{lead})$ output by the adjoint include air-sea feedback. As a result, the convolution in Eq. (7) can erroneously double count air-sea feedback mechanisms.

In order to avoid this problem, we cannot rely on the default configuration of the adjoint. Instead, we have to instruct the algorithmic differentiation software not to take derivatives of the bulk formulae and the surface radiation parameterization code. This approach guarantees that our lagged sensitivity patterns do not include air-sea feedback effects that are already accounted for in the net surface heat flux budget and the net surface freshwater budget. For example, the effect of surface heat fluxes on surface salinity via evaporation is accounted for only in the surface freshwater budget. In addition, the impact of evaporation and precipitation on surface temperature via latent heat fluxes is accounted for only in the surface heat flux budget. However, following our approach, the model's forward trajectory remains the same as in the optimized ECCO state estimate."

Beyond this, the authors show a breakdown based on these 3 forcing mechanisms but there is little text and discussion. Figure 2 has lots of interesting signals that get a superficial discussion at best.

Figure 2 is now Figure 3, and we have added further discussion in lines 371-378 following your recommendation: "For example, salt fluxes, heat fluxes, and wind stress, all, contribute to the positive LSW volume anomalies in the early 2000s. We suggest that the 2008 relative increase in LSW explored by Yashayaev and Loder (2009) can be attributed primarily to surface heat fluxes (Fig. 3a). These attribution results hold true across our three reconstructions. However, all releases of ECCOv4 seem to underestimate the magnitude of the 2008 relative increase in LSW, where the underestimation is most pronounced in release 4. We find that the subsequent 2010-2011 decline in LSW volume is dominated by heat fluxes and wind stress, while the 2012 recovery is attributable to both heat and salt fluxes (Fig. 3a). The well-documented increase in LSW volume after 2015 (Yashayaev and Loder, 2017) seems to be primarily related to wind stress anomalies and, to a smaller extent, surface heat fluxes."
Also, the heat and freshwater spatial flux plots are shown using different units – maybe convert them both to buoyancy flux components, so the reader can more easily compare the magnitude and significance of the terms.

Rescaling the sensitivity to heat and freshwater fluxes in terms of buoyancy fluxes would not be as trivial as it may seem, and here is a brief explanation of the challenges. The rescaling of linear sensitivity patterns would involve thermal expansion and haline contraction coefficients. The issue is determining robustly which local and seasonal coefficients are appropriate to use. The spatial patterns represent seasonal sensitivity of LSW, which occupies intermediate depths in the Labrador Sea. Yet, the fluxes to be rescaled enter the surface over a wide area that extends all the way to the recirculation gyre and the subtropical gyre and in a different season depending on the lead time. If we want to rescale the fluxes into units of buoyancy, we need to use some values for the local coefficients of thermal expansion and haline contraction appropriate for a particular season. It is not so trivial to determine whether the linearized coefficients have to be representative of the deep Labrador Sea in the season of the objective function. Another possibility is that the coefficients have to be representative of the locations along the surface where the fluxes penetrate the ocean in the season corresponding to the lead time. A third option is some weighted average of the equation of state coefficients spanning the pathway between the surface penetration region and the deep Labrador Sea and averaging over a range of seasons.

The case in the present manuscript is particularly challenging because the objective function is evaluated in a region of intensive water mass transformation. This is a major difference

compared to the Kostov et al. (2019) study, where sensitivities of the *subtropical* AMOC to surface fluxes near Greenland were qualitatively rescaled into units of buoyancy. If the reviewer expresses further interest, the first author can explain in further detail why the Kostov et al. (2019) rescaling into buoyancy units was not as challenging.

Adding such a calculation in the present manuscript would require making more assumptions and introducing additional technical explanations about the approximations. The text is already quite loaded with technical explanations because of the nature of the adjoint calculations. We decided against introducing this further complication.

Why rescale to a perturbation order of 10^-8? Why not to order 1, to make the numbers simpler?

We have now clarified this in lines 437-439: "However, we multiply the pattern by a factor of (-10^{-22}), so that the rescaled perturbation is of order $10^{-8}\ m\ s^{-1}$ (or $\sim 10^{-5}\ kg\ m^{-2}\ s^{-1}$), which is comparable to the standard deviation in January surface freshwater fluxes between different years (see Fig. 8b)."

The authors define their sensitivity pattern as a "Traffic Controller". I assume they are hoping this will make the concept easier for the readers to understand. But honestly, I had trouble seeing that acronym and got confused during that discussion. More discussion and a focus to make the definition clear for all readers is needed.

We have now tried to introduce this "nickname" earlier in the text. In lines 17-19, in the abstract, we say: "stress and buoyancy forcing outside the convective interior of the Labrador Sea, at important locations in the North Atlantic Ocean. In particular, patterns of wind and surface density anomalies can act as a "traffic controller" and regulate the North Atlantic Current's (NAC) transport of warm and saline subtropical water masses that are precursors for the formation of LSW."

In addition to all other instances where the "Traffic Controller" is mentioned, we have also added lines 538-540: "This further highlights the importance of the "Traffic Controller" pattern that we identify and its role in driving LSW volume anomalies via alterations to the strength and the pathway of NAC transport."

Following your suggestion, we also refer more to the "Traffic Controller" in the Discussion (lines 579-585): "Some general circulation models show a significant lagged correlation between the upper AMOC limb and LSW, where the former leads the latter in time (Ortega et al., 2017; Li et al., 2019). We are the first to identify this as a causal relationship and an oceanic teleconnection in a state estimate constrained with observations (Forget et al., 2015). Our "Traffic Controller" sensitivity pattern is a geographical fingerprint associated with this oceanic teleconnection that relates flow along the upper AMOC limb to LSW formation and storage in the Labrador Sea. Surface wind stress and density anomalies can act to divert and redirect the transport of warm and saline subtropical water which is necessary for the formation of LSW in the Labrador and Irminger Seas."

Also in the applied perturbation experiments, over exactly what region were the patterns applied?

We have now specified the perturbed region in lines 436-437: “an extended North Atlantic region (north of 20°N, west of 20°E, south of Fram Strait, including marginal seas but excluding the Mediterranean and the Baltic).”

Given the salinification along the Greenland shelf, is there a link to Fram Strait and Arctic outflow/processes?

We have added text in lines 478-480 that describes the importance of temperature and salinity anomalies advected through Denmark Strait:

“Three years after the perturbation, we see the southward transport of anomalously denser (warmer but more saline, Fig. 6b,c) water from the GIN Seas to the Irminger Sea through Denmark Strait.”

However, on the short timescales that we explore, Arctic processes and transport changes through Fram Strait do not play a role.

I also feel there is too much material in the Appendices – too many times the reader is referred to a figure in the appendices, which has relevant material for understanding the main text. For example, the definitions of the regions. I think such information could be included on other figures.

The main text is already busy with figures that convey a lot of information. As a result, we have not moved more graphical information from the Appendices to the main text. However, we have taken note of your concern about the figure panels defining the various regions in Fig. F1 in Appendix F. Following your advice and that of Reviewer 2, we have merged the panels that define regions of interest into one single-panel Fig. F1 where individual regions are distinguished by color. We have also reduced the number of panels in Fig. E1 in Appendix E, where the emphasis now is on the response 36 months after the perturbation.