Environmental Data Science (2024), **xx**:xx 1–**??**

CAMBRIDGE
UNIVERSITY PRESS

APPLICATION PAPER

Supplementary Materials for: Unsupervised clustering identifies thermohaline staircases in the Canada Basin of the Arctic Ocean

Mikhail G. Schee[1], Erica Rosenblum[1,2], Jonathan M. Lilly[3] and Nicolas Grisouard[1]

[1]Department of Physics, University of Toronto, Toronto, M5S 1A7, Ontario, Canada, E-mail: mikhail.schee@mail.utoronto.ca.
[2]Centre for Earth Observation Science, University of Manitoba, Winnipeg, R3T 2M6, Manitoba, Canada.
E-mail: erica.rosenblum@utoronto.ca.
[3]Planetary Science Institute, Tucson, 85719, Arizona, USA. E-mail: jmlilly@psi.edu.

Keywords: Arctic Ocean, clustering algorithm, machine learning, thermohaline staircases, ice-tethered profilers

S. Supplementary materials

S.1. Practical salinity S_P vs. absolute salinity S_A

TEOS-10 recommends storing values of practical salinity S_P in databases in order to maintain continuity with older records and also because S_P is practically a measured quantity [5]. In contrast, absolute salinity S_A is a calculated variable defined as "the mass fraction of dissolved non-H_2O material in a seawater sample at its temperature and pressure," and is recommended for use in publications [3]. For this study, we ultimately chose to work with S_P because the two studies to which we make direct comparisons, T08 and L22, used S_P. Specifically, when we define the ranges of S_P to filter to before running the clustering algorithm, we were able to directly take the values of S_P from T08 and L22 without having to first convert to S_A.

To investigate the difference between using S_P vs. S_A on our results, we recreated Figure 2, but replacing S_P with S_A on all the axes, as shown in Figure S.1. We still filtered the ITP2 data to be in the practical salinity range S_P 34.05–34.75 g/kg to make sure we used the same data points. However, then we plotted those data points in Θ'–S_A space in panel (c) to run the clustering algorithm using the same $m_{pts} = 170$. This resulted in 38 clusters (as opposed to 36) with a DBCV score of 0.2177 (as opposed to 0.3034). In panel (d), the values of R_L displayed made minor changes of (from left to right): +0.39, -0.04, -0.03, -0.03, -0.01, +0.14, and +0.07.

We also reproduced Figure 5(a,b) in Figure S.2 using the clustering shown in Figure S.1. The two outliers in IR_{S_A} in Figure S.2(a) do not correspond to outliers in Figure 5(a) because these have very high IR_{S_A}, over 20, which skews the mean as used when calculating the z-score. The dark green star and the red triangle in Figure S.2(a) correspond to the two outliers in Figure 5(a) where they are marked by an orange 4-pointed star and a green "×". These pairs of points have very similar values of IR_{S_A} and IR_{S_P}, respectively.

In Figure S.2(b), we see that there are two outliers in R_L circled, both of which correspond to outliers found when using S_P. The purple star in Figure S.2(b) corresponds to the orange left half circle in Figure 5(b), while the dark green "Y" in Figure S.2(b) corresponds to the teal "×" in Figure 5(b). A notable difference is that the purple star outlier in Figure 5(b) (which spans S_P = 34.233–34.261 in Figure 2(c) encompassing what should be two distinct clusters) is no longer present in Figure S.2(b). Comparing Figures 2(c) and S.1(c), we see that it is now replaced by two clusters, the dark green left half circle and the red right half circle, neither of which are outliers in R_L.

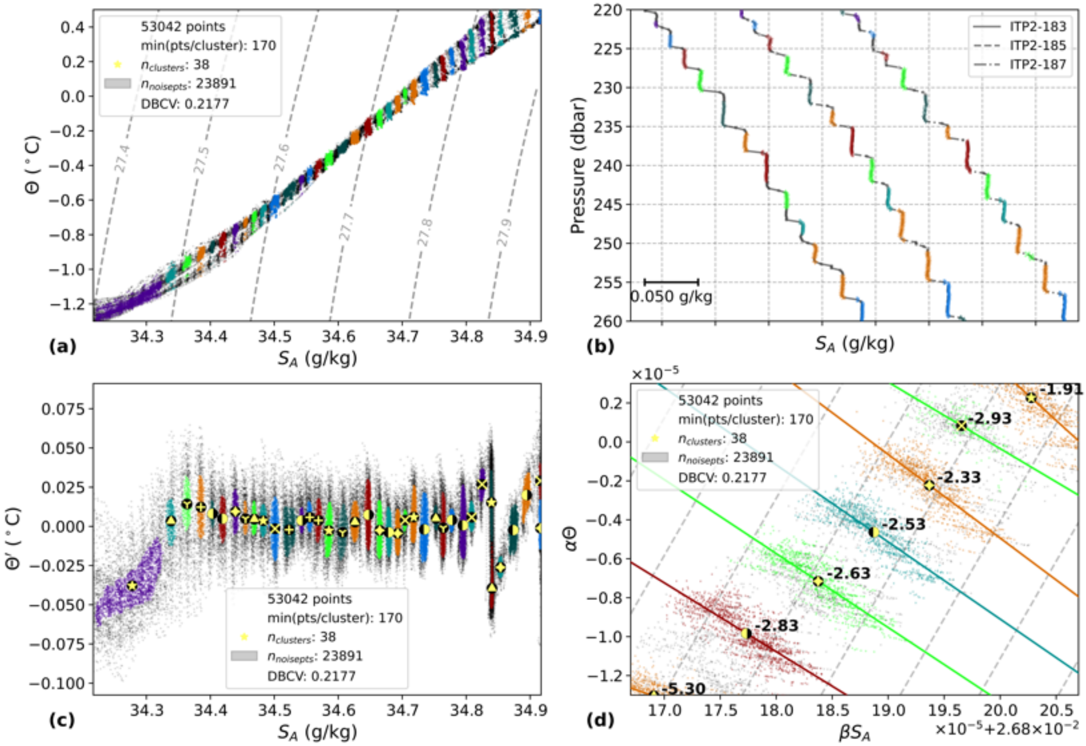

Figure S.1. Results from the clustering algorithm with m_{pts} = 170 and ℓ = 100 dbar run on 53042 data points in the S_P range 34.05–34.75 g/kg from all up-going ITP2 profiles. This is the same as Figure 2, but the clustering algorithm was run on absolute salinity S_A instead of practical salinity S_P. (a) The data in Θ–S_A space with dashed lines of constant potential density anomaly (kg m^{-3}) referenced to the surface. (b) Profiles 183, 185, and 187 from ITP2 in a limited pressure range to show detail. Each profile is offset in S_A for clarity. (c) The spatial arrangement used as input for the algorithm where the gray points are noise and each color-marker combination indicates a cluster. The same color-marker combinations are used in each panel and the markers in panels (c) and (d) are at the cluster average for each axis. (d) A subset of the data in $\alpha\Theta$–βS_A space with the linear regression line and inverse slope (R_L) noted for each individual cluster and with dashed lines of slope $\alpha\Theta/\beta S_A = 1$.

We also see that the 2nd-degree polynomial fit in Figure S.2(b) is similar to that in Figure 5(b). We conclude that the difference between using S_A vs. S_P does not significantly affect our results.

S.2. Varying the value of ℓ

We define the local anomaly of a temperature profile Θ′ as the original profile minus a version of that profile smoothed with a moving average. Subtracting a smoothed profile from the original profile in this way gives temperature differences that are centered around zero and leads to more continuous clusters (compare Figures 2(a) and 2(c)), allowing HDBSCAN to group these points more accurately.

When calculating Θ′ with Equation 3.1, we use a moving average with window size ℓ, so the resulting profile has dead regions $\ell/2$ in size with no data at the top and bottom. Therefore, ℓ cannot be larger than twice the distance available in the profile, either above or below the pressure range to be analyzed, whichever is smaller. As ℓ increases, the moving average profile gets flatter, meaning Θ′ becomes closer to the original profile, just shifted in temperature space. Choosing a very small ℓ will give a moving average temperature profile that will closely match the original temperature profile, meaning that the Θ′

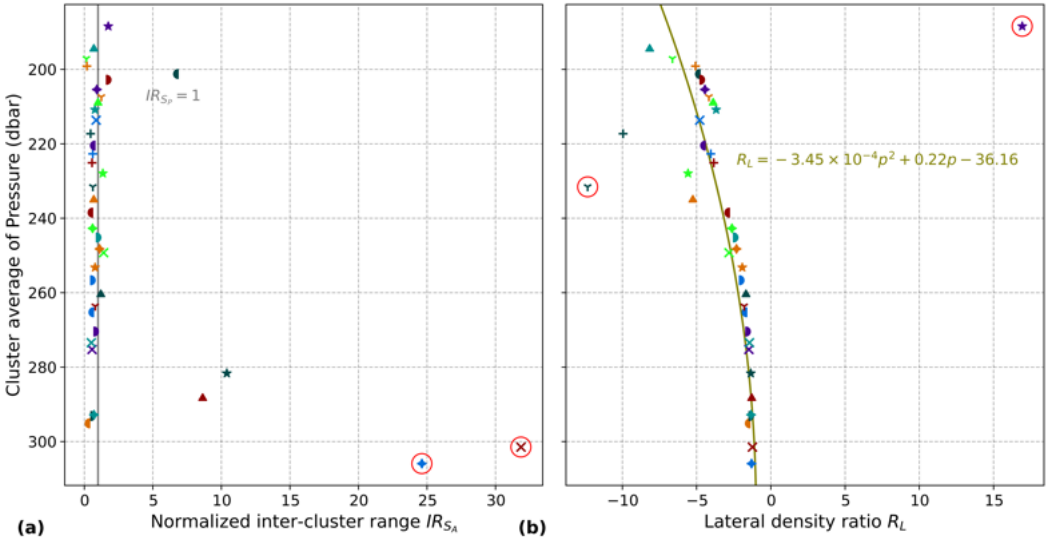

Figure S.2. *The value of each cluster's normalized inter-cluster range for salinity* IR_{S_A} *in (a) and the lateral density ratio* R_L *in (b) as a function of the cluster's average pressure for ITP2. This is the same as Figure 5(a,b), but the clustering algorithm was run on absolute salinity* S_A *instead of practical salinity* S_P*. The colors and markers are the same as the clustering shown in Figure S.1. Markers circled in red indicate outliers with a z-score greater than* 2.

profile will be very close to zero. This eliminates any effect temperature might have on the clustering results.

In Figure S.3, we show 4 pairs of profiles, restricted to the 200-400 dbar pressure range to show detail. All profiles are number 185 from ITP2, the example profile featured in Figure 2(b). Each panel shows both the original profile and the moving average profile with the values of ℓ: 10, 50, 100, and 150 dbar. The original and moving average profiles are slightly offset for clarity. When $\ell = 10$ dbar, the moving average profile still clearly contains some of the larger stair steps. When $\ell = 150$ dbar, almost all of the features in the original profile are gone.

We noted a feature in 2(c) where the cluster average Θ' increases starting at $S_P \approx 34.63$ g/kg until it jumps sharply to negative Θ' values around $S_P \approx 34.67$ g/kg, then increases again until $S_P \approx 34.72$ g/kg. This zig-zag pattern is due to the choice of ℓ and the presence of the AW subsurface temperature maximum in Θ profiles. In Figure S.4, we show the ITP2 data used in the study in Θ'–S_P space for the same 4 values of ℓ as in Figure S.3. This clearly illustrates that the zig-zag pattern is not present for low ℓ and becomes more pronounced as ℓ increases.

For each panel in Figure S.4, we ran a parameter sweep across m_{pts} similar to that shown in Figure 3(b), selecting the value of m_{pts} which gave the highest DBCV, as shown in the legends. If we always choose ℓ to be approximately twenty times the typical layer thickness, then the panels in Figure S.4 represent clusterings for thickness estimates of 0.5 dbar, 2.5 dbar, 5 dbar, and 7.5 dbar. We conclude that the results are not significantly sensitive to this typical layer thickness estimate as the number of clusters and their positions along the S_P axis are generally consistent across the four values of ℓ, especially between $S_P \approx 34.18$ g/kg and $S_P \approx 34.62$ g/kg. We find the most reasonable clusterings occur when ℓ is chosen to be small enough that the features outside the pressure range we analyze do not significantly affect the moving average, yet large enough that the stair steps are completely smoothed out.

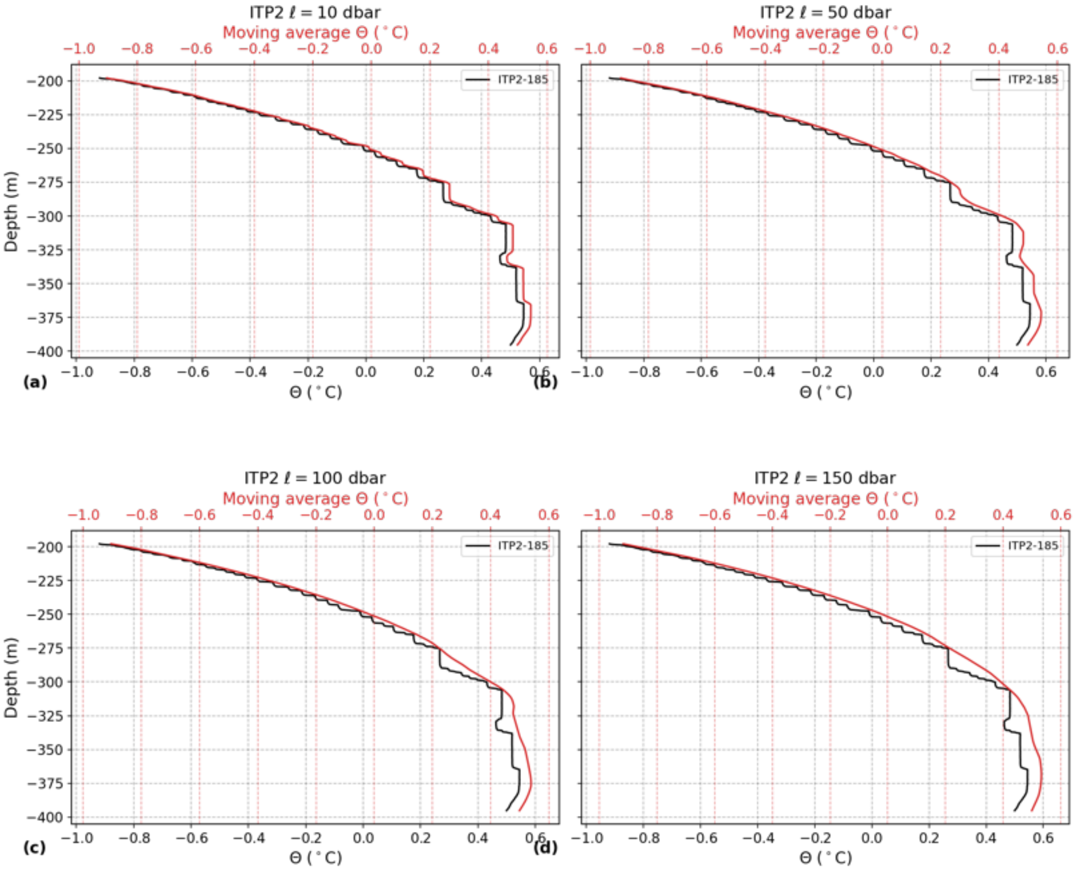

Figure S.3. *The original profile 185 from ITP2 and the smoothed version of that profile for 4 different values of ℓ: (a) 10 dbar, (b) 50 dbar, (c) 100 dbar, and (d) 150 dbar. The Θ axes on each panel are offset slightly to show both lines.*

S.3. Optional parameters for HDBSCAN

While m_{pts} has the most significant effect on the clustering results, there are several other parameters which could be adjusted when using the “hdbscan” Python package. In each case, we chose to follow the recommendation of the authors of the algorithm [2, 7]. The default distance metric is Euclidean and, as we cluster in a two dimensional space of continuous numerical variables, we maintain this selection. For the cluster selection method, we chose Leaf over Excess of Mass to avoid the results collapsing into fewer, larger clusters. HDBSCAN can be combined with DBSCAN by setting a threshold of cluster selection $\varepsilon_{clSelect}$ to explicitly group clusters a certain distance apart. However, we chose the default behavior of HDBSCAN which automatically identifies the values of ε which give the most stable clusters. HDBSCAN considers a neighborhood to be dense when m_{pts} points are within a certain distance ε. While m_{pts}, known as the minimum samples, can be independently set higher to more liberally include fringe points in clusters or lower to more conservatively classify them as noise, we follow the default recommendation to set this equal to the minimum points per cluster m_{clSize}.

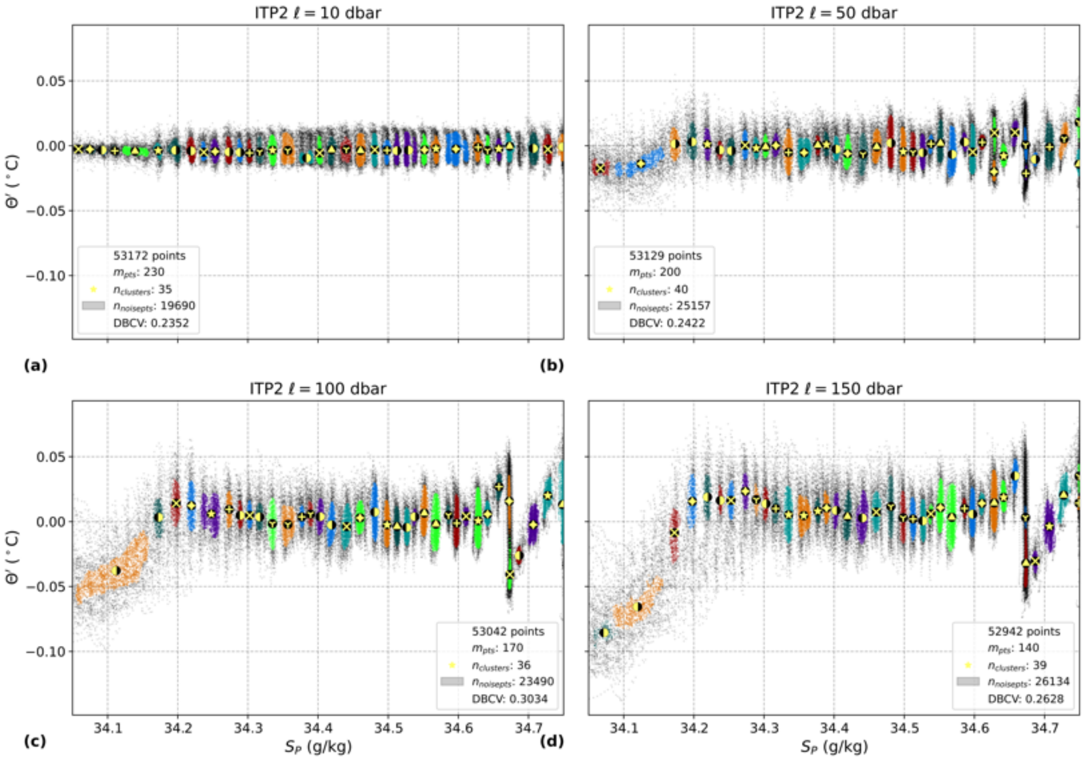

Figure S.4. *Clustered data from ITP2 in the salinity range 34.05—34.75 g/kg plotted in Θ'–S_P space for 4 different values of ℓ: (a) 10 dbar, (b) 50 dbar, (c) 100 dbar (the same as in Figure 2(c)), and (d) 150 dbar, where the gray points are noise and each color-marker combination indicates a cluster.*

S.4. Clusters from ITP3

In Figure S.5, we create a figure similar to that of Figure 2, but for ITP3. In Figure S.6, we create a figure similar to that of Figure 3(a), but for ITP3. Note that the run with $m_{pts} = 710$ had a DBCV score of 0.3885, which is slightly higher than the run at $m_{pts} = 580$ with a DBCV score of 0.3862. However, because the differences in DBCV were not great, we decided to use $m_{pts} = 580$ to reduce computation time as the clustering algorithm takes longer to run with higher values of m_{pts}.

S.5. Clusters across time for ITP2

In Figure S.7, we create a figure similar to that of Figure 4, but for ITP2.

S.6. Differences from T08

Figure 6 from Timmermans et al. [8] shows clusters in $\alpha\Theta - \beta S_P$ space with panel (a) showing five values of R_L for ITP2, which range from -3.5 to -3.0, and panel (b) showing 12 values of R_L for ITP5, which range from -3.9 to -2.4. It is unclear whether their reported $R_L = -3.7 \pm 0.9$ applies to data from ITP2, ITP5, both, or all six ITPs they included in their study. It is not explicitly stated whether that value applies to just the clusters shown in that figure or it includes the dozens of clusters they identified earlier in their study. T08 never explicitly state that they did not find a trend in R_L with respect to depth. However, given that T08 reported one overall value of R_L and that Bebieva and Timmermans [1] (which

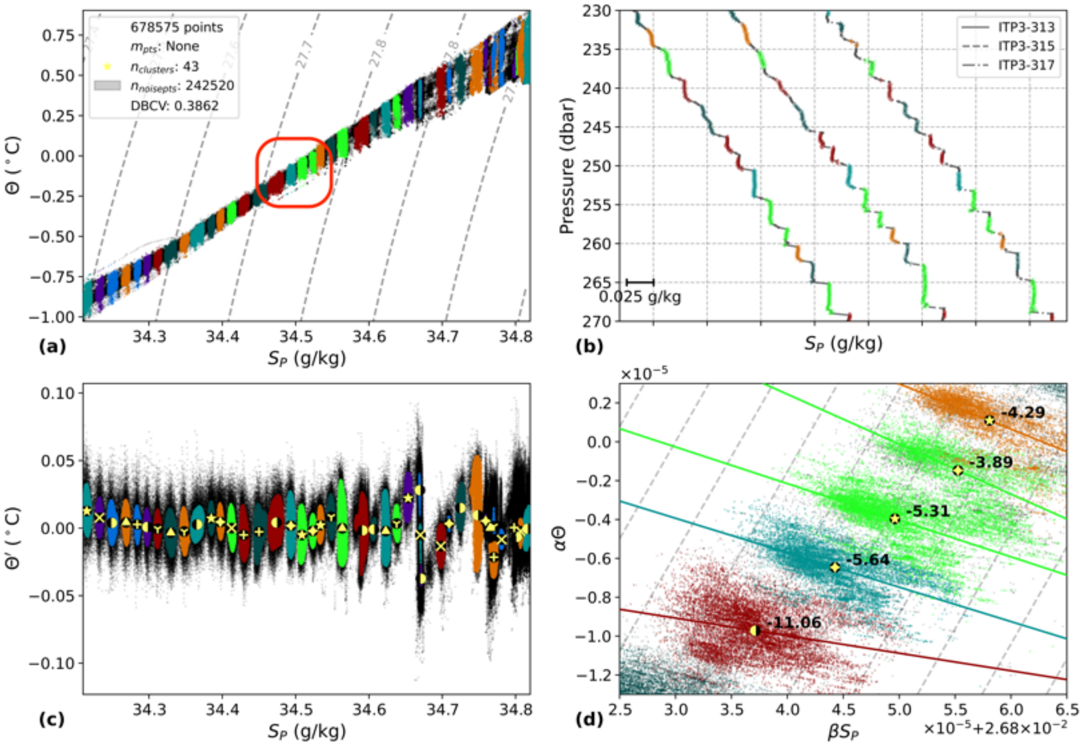

Figure S.5. *Results from the clustering algorithm with m_{pts} = 580 and ℓ = 100 dbar run on 678575 data points in the S_P range 34.21–34.82 g/kg from all up-going ITP3 profiles. (a) The data in Θ–S_P space with dashed lines of constant potential density anomaly (kg m^{-3}) referenced to the surface. (b) Profiles 313, 315, and 317 from ITP3 in a limited pressure range to show detail. Each profile is offset in S_P for clarity. (c) The spatial arrangement used as input for the algorithm where the gray points are noise and each color-marker combination indicates a cluster. The same color-marker combinations are used in each panel and the markers in panels (c) and (d) are at the cluster average for each axis. (d) A subset of the data in αΘ–βS_P space with the linear regression line and inverse slope (R_L) noted for each individual cluster and with dashed lines of slope $\alpha\Theta/\beta S_A = 1$.*

extended the work of T08) show a constant value of R_L in the depth range we consider, we assume T08 also found a constant value of R_L.

There are several factors that could explain why our results for R_L differ from those of T08. The fact that they used potential temperature θ while we worked with conservative temperature Θ is not relevant, since when we repeat our calculations using θ instead of Θ, the values of R_L we find are negligibly different. It is unclear whether T08 calculated the values of α and β for each point or for each layer. We calculated α and β for each point in each profile individually using functions from the Gibbs Seawater (GSW) Oceanographic Toolbox, the Python implementation of TEOS-10 [5], while T08 used the McDougall [4] algorithm coded by Morgan [6] as GSW was not yet released. This difference in the calculation of α and β may conceivably explain why the clusters we show in Figure 2(d) are tilted several degrees compared to those shown in Figure 6(a) from T08, and why the values of R_L found here are all less negative than corresponding values in Figure 6(a) of T08. The difference in calculating α and β may also account for the shift in the ranges of βS_P between the two figures. Both plots show a span in βS_P of 4.0×10^{-5}, but our plot ranges from 2.6838×10^{-2} to 2.6878×10^{-2}, while theirs ranges from 2.7002×10^{-2} to 2.7042×10^{-2}. Because the clustering algorithm only considers Θ′ and S_P, any difference in calculating α and β would not affect the clusters we found, only the values of R_L.

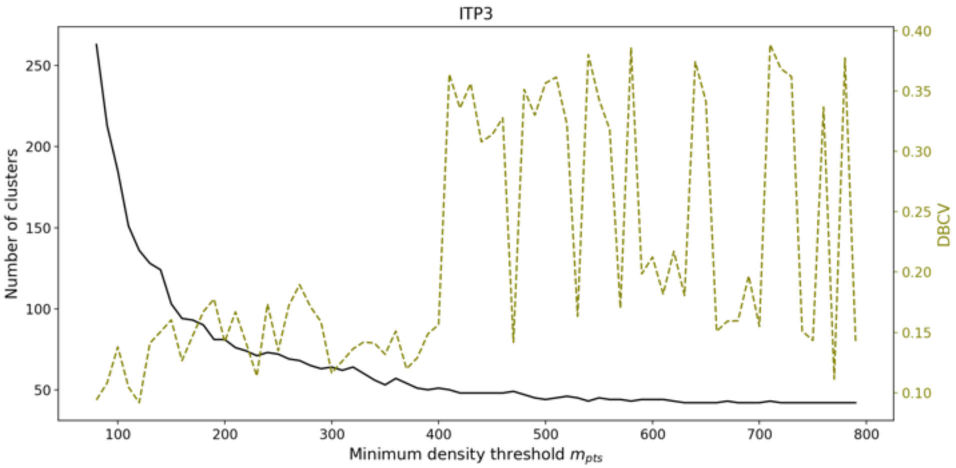

Figure S.6. *A parameter sweep showing the number of clusters found (solid lines) and DBCV (dashed lines) in ITP3 as a function of 72 different values of* m_{pts} *with* $\ell = 100$ *dbar.*

S.7. Differences from L22

L22 used 758 profiles from ITP3 while we used all 766 available up-going profiles. The reasons 8 profiles were not used was not stated, nor could we determine exactly which profiles they are. Based on the gaps in Figure 3 of L22, we believe that the 8 missing up-going profiles are from July 2006. We do not, however, believe that this difference of 8 profiles accounts in any significant way for the differences between our results and those of L22.

Instead, we believe the reason our results differ from those of L22 has to do with the difference in the methods used to identify layers. Because the method used by L22 to delineate different layers only considered salinity, they would not separate layers with temperature inversions. Remnant intrusions are homogeneous in salinity but, compared to stair steps, are inverted in temperature and have different patterns of heat and salt flux. They tend to be present near the bottom of the staircase, which is indeed where we find the largest differences between our results and those of L22. A remnant intrusion's pattern of a section of warmer water above a section of colder water is distinct enough for the clustering algorithm to find two clusters. However, if only salinity is considered, as L22 did, it would appear to be one regular staircase layer. Whether a remnant intrusion represents one or multiple layers is subjective, and we do not claim that one method is superior to the other. Here, in fact, it is the disagreement between these two methods is arguably the most interesting point, because it indicates that a specific physical process is operating.

While we agree with L22 that salinity is the best measured quantity by which to identify the layers of a staircase, it is still important to consider temperature. Another topic of future study would be to adapt this method to automatically distinguish between well-mixed layers and intrusions.

References

[1] Bebieva, Y. & Timmermans, M. L. (2019). Double-Diffusive Layering in the Canada Basin: An Explanation of Along-Layer Temperature and Salinity Gradients. *Journal of Geophysical Research: Oceans*, *124*(1), 723–735, https://doi.org/10.1029/2018JC014368.

[2] Campello, R. J. G. B., Moulavi, D., & Sander, J. (2013). Density-Based Clustering Based on Hierarchical Density Estimates. *Lecture Notes in Computer Science*, *7819*, 160–172, https://doi.org/10.1007/978-3-642-37456-2_14.

[3] IOC, SCOR, & IAPSO (2010). The international thermodynamic equation of seawater – 2010 : Calculation and use of thermodynamic properties, volume 56. UNESCO.

[4] McDougall, T. J. (1987). Neutral Surfaces. *Journal of Physical Oceanography*, *17*(11), 1950–1964, https://doi.org/10.1175/1520-0485(1987)017<1950:ns>2.0.co;2.

[5] McDougall, T. J. & Barker, P. M. (2011). Getting started with TEOS-10 and the Gibbs Seawater (GSW) Oceanographic Toolbox. 3.06.12 edition www.TEOS-10.org.

[6] Morgan, P. P. (1994). SEAWATER: A Library of MATLAB® Computational Routines for the Properties of Sea Water: Version 1.2. Technical report, CSIRO Marine Laboratories, Hobart, Tasmania http://hdl.handle.net/102.100.100/239771?index=1.

[7] Moulavi, D., Jaskowiak, P. A., Campello, R. J. G. B., Zimek, A., & Sander, J. (2014). Density-Based Clustering Validation. In *Proceedings of the 2014 SIAM International Conference on Data Mining*, volume 2 (pp. 839–847). Philadelphia, PA: Society for Industrial and Applied Mathematics https://epubs.siam.org/doi/10.1137/1.9781611973440.96.

[8] Timmermans, M.-L., Toole, J., Krishfield, R., & Winsor, P. (2008). Ice-Tethered Profiler observations of the double-diffusive staircase in the Canada Basin thermocline. *Journal of Geophysical Research*, *113*, 1–10, https://doi.org/10.1029/2008jc004829.

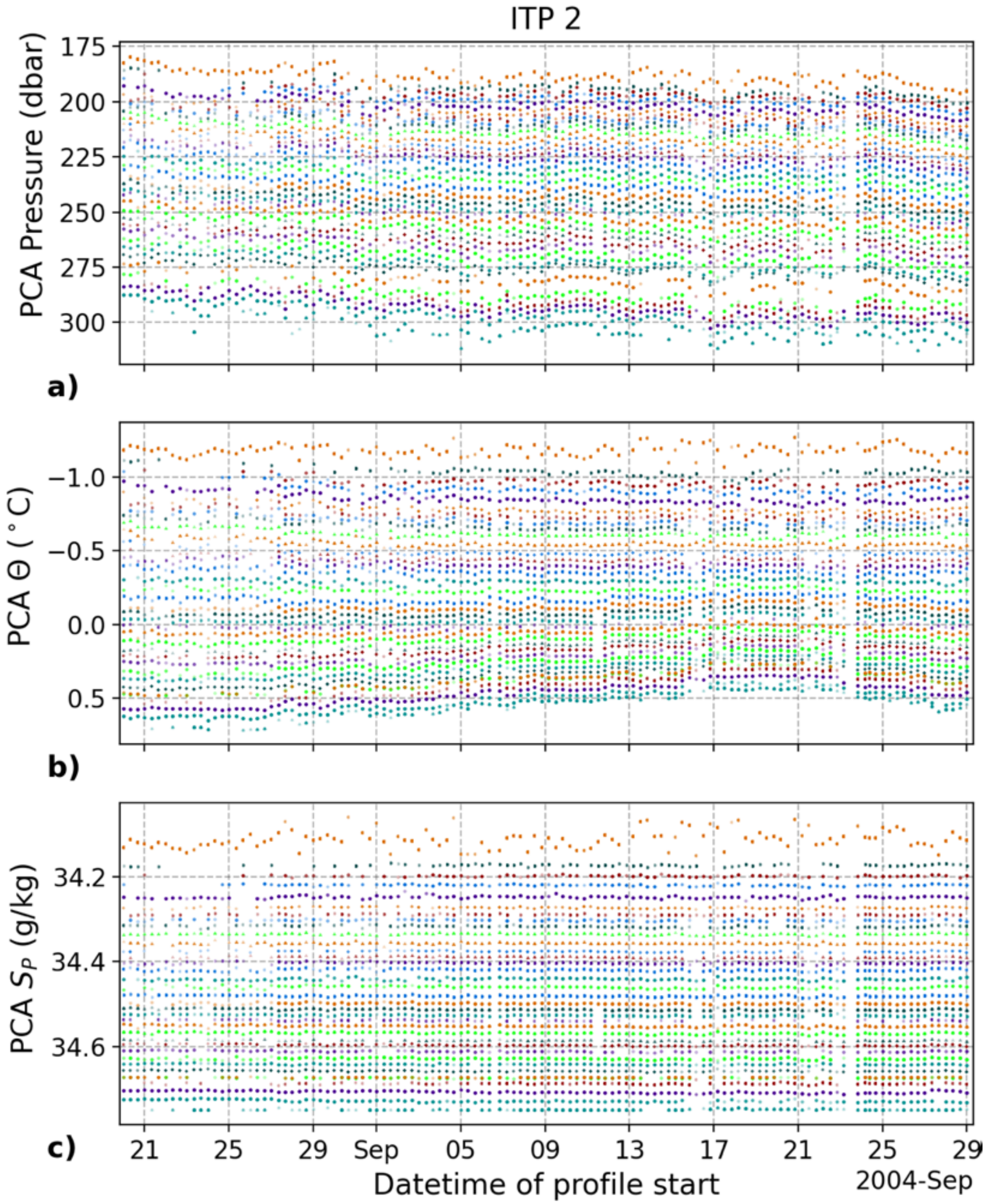

Figure S.7. *The average (a) pressure, (b) Θ, and (c) S_P for the points within each cluster for each profile (profile cluster average, PCA) across time. The clustering algorithm was run with m_{pts} = 170 and ℓ = 100 dbar on 53042 data points in the salinity range 34.05–34.75 g/kg from all up-going ITP2 profiles.*