A clustering approach to determine biophysical provinces and physical drivers of productivity dynamics in a complex coastal sea

Tereza Jarnikova[1], Elise M. Olson[1], Susan E. Allen[1], Debby Ianson[2,1], and Karyn D. Suchy[1]

[1]Department of Earth, Ocean, and Atmospheric Sciences, University of British Columbia, Vancouver, BC, Canada
[2]Institute of Ocean Sciences, Fisheries and Oceans Canada, Sidney, BC, Canada

Correspondence: Tereza Jarnikova (T.Jarnikova@uea.ac.uk)

Abstract. The balance between ocean mixing and stratification influences primary productivity through light limitation and nutrient supply in the euphotic ocean. Here, we apply a hierarchical clustering algorithm (Ward's method) to four factors relating to stratification (wind energy, freshwater index, watercolumn-averaged vertical eddy diffusivity, and halocline depth), as well as to depth-integrated phytoplankton biomass, extracted from a biophysical ocean model of the Salish Sea. We then assess these factors for spatial co-occurrence. Running the clustering algorithm on four years of model output, we identify distinct regions of the model domain that exhibit contrasting wind and freshwater input dynamics, as well as regions of varying watercolumn-averaged vertical eddy diffusivity and halocline depth regimes. The spatial regionalizations in physical variables are similar in all four analyzed years. We also find distinct interannually consistent biological zones. In the Northern Strait of Georgia and Juan de Fuca Strait, a deeper winter halocline and episodic summer mixing coincide with higher summer diatom abundance, while in the Fraser River stratified Central Strait of Georgia, shallower haloclines and stronger summer stratification coincide with summer flagellate abundance. Cluster-based model results and evaluation suggest that the Juan de Fuca Strait supports more biomass than previously thought. Our approach elucidates probable physical mechanisms controlling phytoplankton abundance and composition. It also demonstrates a simple, powerful technique for finding structure in large datasets and determining boundaries of biophysical provinces.

1 Introduction

Marine phytoplankton form the basis of the oceanographic food web and are responsible for approximately half of global carbon fixation (Field et al. (1998)). To predict changes in global ecosystem functioning, it is necessary to understand the underlying controls on marine productivity. Primary productivity in the near-surface ocean is controlled by the availability of macro- and micro-nutrients and light, as well as temperature, which are in turn controlled by the interplay of stratifying processes and sources of mixing.

The breakdown of the surface ocean stratified layer may reduce the availability of light for phytoplankton, inhibiting growth (e.g. Sverdrup (1953)), or contrastingly bring nutrients from deeper waters to nutrient-depleted surface waters, thus stimulating growth. The interplay of different stratification regimes exerts control on the structure of ocean ecosystems (e.g. Legendre

(1981)), and changes in regime have been linked to shifts in phytoplankton community composition (e.g. Huisman et al. (2004)).

The importance of phytoplankton in biogeochemical cycling, as well as their position at the base of the food web and impact on higher trophic levels, globally motivates the study of phytoplankton distribution and dynamics. Coastal regions are more productive than the open ocean (e.g. Longhurst et al. (1995)). Simultaneously, these regions typically have more complex mixing, circulation, and stratification dynamics than the open ocean, making resolution of phytoplankton biomass patterns difficult. Finally, because both ocean stratification patterns and phytoplankton biomass dynamics may be expected to shift under anthropogenic climate change (Richardson (2008)), there exists a need to characterize their dynamic structure and identify key drivers.

1.1 Oceanographic Setting

The Salish Sea is a semi-enclosed fjordlike estuary on the British Columbia coast, composed of the Strait of Georgia (SoG), Juan de Fuca Strait (JdF), and Puget Sound (Fig. 1). The SoG is connected to the open ocean by Juan de Fuca Strait to the south and Johnstone Strait to the north, with Juan de Fuca Strait serving as the site of primary seawater exchange with the open ocean (Khangaonkar et al. (2017)). The Salish Sea receives freshwater input from over 200 rivers, but the primary freshwater source is the nival-glacial Fraser River (Pike et al. (2010)), which drives high salinity-induced stratification in the CSoG and a strong estuarine exchange (Giddings and MacCready (2017)). Salinity stratification is opposed by wind and tidal action. Strong winds in the fall and winter months lead to mixing of surface and intermediate water masses. The SoG contains two deep basins (North and Central), with the Fraser River plume sitting on top of the Central basin. Deep SoG water is relatively unmixed, except during deep water renewal events (Masson (2002)).

This coastal ocean is a region of ecological and cultural importance, providing habitat to important megafauna, including the Southern Resident killer whales (*Orcinus orca*) and the local salmon populations. The ongoing significant decline of the local Coho and Chinook salmon (Preikshot et al. (2013)) has been implicated as a factor in the low reproductive success of the killer whale populations (Wasser et al. (2017)), which depend on these salmon as a food source. The health of fish populations in the Pacific Northwest has been linked to spring bloom timing and phytoplankton abundance (e.g. Malick et al. (2015); Boldt et al. (2019)). Thus, potential population declines in upper trophic levels further motivate the understanding of factors controlling the base of the food web.

The physical environment of the Salish Sea is well known, with functionally-distinct physical-oceanographic regions (Thomson (1981); LeBlond (1983); Pawlowicz et al. (2020)). An ongoing subject of interest in this coastal sea is the relationship between known physical, and presumed ecological, regions. Three prominent parts of the Salish Sea - Juan de Fuca Strait (JdF), the Northern Strait of Georgia (NSoG), and the Central SoG (CSoG) have been defined by distinct stratification regimes and watermass characteristics, and available biological observations and model results (e.g. Masson and Peña (2009); Suchy et al. (2019); Peña et al. (2016)) are typically discussed in the context of these differing physical environments. However, in situ sampling of phytoplankton biomass remains relatively sparse and episodic, and may not capture inherently dynamic phytoplankton biomass fluctuations, while remote sensing approaches can provide only surface chlorophyll concentrations. Here,

we aim to use an unsupervised cluster analysis of a well-resolved submesoscale mechanistic biophysical model to consider the linkages between the regional physical oceanography of the system and its phytoplankton biomass dynamics.

1.2 Application of clustering methods to a modelling framework

Clustering methods have demonstrated utility in identifying underlying structures in large observational datasets and are commonly used in ecological and biological observational studies. In recent years, the application of clustering to physical and biogeochemical ocean models has become more common (e.g. Sonnewald et al. (2020); Follows et al. (2007); Sun et al. (2021)), though these approaches are not yet in widespread use. The quantity of data motivates the use of clustering methods in a modelling context - even in our relatively spatially limited sub-mesoscale resolution model, one year of output of a single variable at hourly resolution is quite sizeable ($\sim$ 60 GB and $\sim 3 \times 10^{10}$ individual values); the output of global circulation models is considerably larger. Well-tuned, high-resolution numerical models of complex natural systems are uniquely poised to provide insight regarding physical oceanographic regimes and overarching patterns, especially in diverse regimes where sampling efforts are sparse and often seasonally biased to fair weather. However, interpreting (even visualising) large volumes of data poses a unique challenge; common approaches, such as monthly-averaged map snapshots, may represent an oversimplification and fail to show the patterns present in the underlying system. By extracting small-data key metrics throughout the model domain, we reduce the size of the problem we are considering while keeping the key characteristics of the system that we are studying. We can then cluster these metrics in the hopes of revealing discrete dynamical regimes in complex regions. Furthermore, the clustering method is an objective classification of the system in the sense that it makes no prior assumptions about the locations of any oceanographic features that it finds.

Here our main goal is to investigate how physical dynamics in the Salish Sea objectively define regions of distinct phytoplankton biomass and functional group composition. We extract model-available proxies for four separate factors related to watercolumn stratification: wind energy, freshwater index, watercolumn-averaged vertical eddy diffusivity, and halocline depth, and one indicator of primary productivity (depth-integrated phytoplankton biomass separated by functional group). We then cluster each factor individually in order to discuss the three major regions of the Salish Sea in the context of the spatial patterns in the yearly signals of these factors, as well as to consider their interannual variability. We finally compare spatial patterns in stratification factors to spatial patterns in phytoplankton biomass and discuss possible linkages between the two.

2 Methods

2.1 The SalishSeaCast biophysical model

We use SalishSeaCast, a regional oceanographic model developed for the Salish Sea (Soontiens et al. (2016); Soontiens and Allen (2017)) using version 3.6 of the NEMO regional ocean modeling engine (Madec et al. (2017)). The physical model solves the Reynolds-averaged Navier-Stokes equations on an Arakawa-C grid, with a 2 second barotropic timestep, a 2 second vertical

advection timestep, and a 40 second baroclinic timestep. Major physical model modifications since first implementation are summarized in Olson et al. (2020).

The model domain (Fig. 1) is 898 (v) by 398 (u) horizontal cells with approximately 500 m horizontal resolution and 40 vertical z -layers ranging from 1 m resolution at the surface to 27 m resolution at the bottom. SalishSeaCast has two open boundaries, at JdF and Johnstone Strait, which are forced with eight tidal constituents and sea surface height predictions from NOAA's storm surge forecast at Neah Bay, in JdF near the seaward entrance. The model is forced with over 150 rivers; the Fraser River runoff is taken from the Environment and Climate Change Canada flow gauge at Hope, BC, and the remaining rivers, as well as the Fraser River downstream of Hope, are forced by a monthly climatology (Morrison et al. (2012)). Atmospheric forcing, including winds and solar radiation, is derived from the High Resolution Deterministic Prediction System (HRDPS), a nested 2.5 km resolution operational atmospheric model (Milbrandt et al. (2016)). The HRDPS model output is too coarse to accurately resolve atmospheric conditions in the northern inlets, but its wind fields have shown good agreement with observations throughout the Strait of Georgia (Moore-Maley and Allen (2022)). Coupled to the physical model is a NPZD-type biological model (SMELT - Salish Sea Lower Trophic Ecosystem Model, Olson et al. (2020)), which is described in summary below.

The SMELT biological model represents the transfer of matter, using nitrogen as currency, between three classes of primary producers (diatoms, small flagellates, and the ciliate mixotroph *M. rubrum*), three classes of nutrients (nitrate, ammonia, and silicic acid), three classes of detritus (particulate and dissolved organic nitrogen, and biogenic silica) and one class of microzooplankton, with mesozooplankton grazing as a closure term. The growth rate of all three primary producer classes depends on the availability of nutrients, light, and on temperature. The diatom class is assigned the highest maximum growth rate and the highest optimal light level and is the only class to take up dissolved silica – in the gleaner-opportunist framework, we consider it an opportunist class (Grover (1990); Grover et al. (1997)). Small flagellates (representing phytoplankton groups such as cryptophytes) have the lowest maximum growth rate while competing better at low nitrogen levels, low light, and higher temperature. Small flagellates have the lowest minimum nutrient requirement, and we consider them the gleaner class in the gleaner-opportunist framework. The mixotroph *M. rubrum* has intermediate growth parameters while grazing on the flagellate class in addition to photosynthesizing. Details of phytoplankton growth rate and nutrient and light level preference are available in Olson et al. (2020). A summary of minor updates to the model tuning since publication in Olson et al. (2020) is provided in Appendix B.

SalishSeaCast has been run operationally since 2014, and results from a 2013 to 2021 hindcast are available at https://salishsea.eos.ubc.ca/erddap/index.html. The entire model system, including run environment, is documented at https://salishsea-meopar-docs.readthedocs.io. In Appendix A we provide an evaluation of the model salinity, temperature, nitrate, dissolved silica, and chlorophyll against available observations for the years and model version analyzed, separated according to the major clusters found (Fig. A1-A2). In summary, the model shows consistently high skill in across all clusters (Tables A1-A2), with Willmott skill scores for temperature and salinity ranging from 0.957-0.971 and 0.959-0.971 respectively across the clusters, while comparisons with log-transformed total chlorophyll data yield scores of 0.599-0.712.

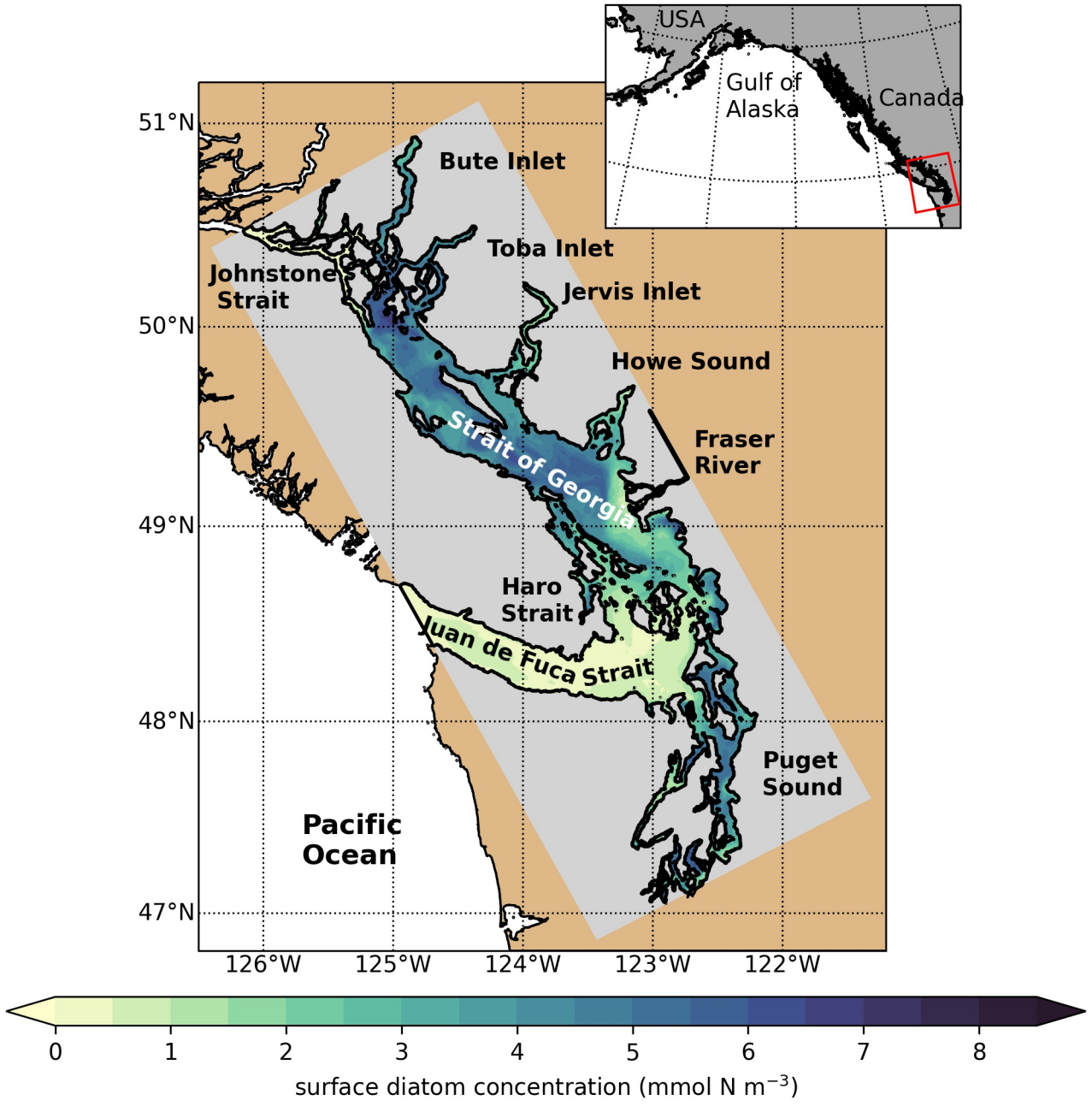

Figure 1. SalishSeaCast model domain coloured by one day of surface diatom concentration (April 1, 2016), highlighting major geographic subregions and features. The Strait of Georgia is often subdivided into the Central Strait of Georgia (CSoG) and Northern Strait of Georgia (NSoG).

2.2 Stations and clustering signals

We analyzed four years of daily output from a hindcast of SalishSeaCast (2013-2016), using an unsupervised clustering algorithm (Ward's Euclidean Distance Method, see section 2.3). We developed model-available year-long timeseries proxies ("signals") for four different factors relating to stratification and mixing activity (wind strength, freshwater influx, vertical eddy diffusivity, and halocline depth) and one for an indicator of biological productivity (total depth-integrated biomass of three phytoplankton functional groups from the model's NPZD module). These signals were extracted for each year at each of 571 model "stations" spaced 10 model grid points apart ($\sim$5 kilometers, Fig. 2). This spacing was chosen as a compromise between resolution and computing time, and we believe it well represents the different regions of the Salish Sea while being computationally manageable.

Several possible clusterings resulting from our analysis were visualized and compared for major differences (see Section 2.3). Results show the typical cluster structure for all four years for each individual factor (Section 3), while an example visualization of all possible clusterings of one year of one of the variables is available in Fig. C1. Here, we describe the signals.

Wind strength

The wind forcing used (HRDPS, see model description) has 2.5 km spatial resolution and hourly temporal resolution, and is used operationally by Environment Canada in the Canadian Pacific region. The skill of the HRDPS wind product in this region when compared to local meteorological stations has been evaluated by a previous study and accurately reproduces the climatology of observed wind magnitudes and directions (Moore-Maley and Allen (2022)). We first interpolate this product onto the model grid and then extract hourly windspeed. Here we are interested in the impact of wind on mixing of the watercolumn. Therefore, because wind energy available for mixing scales with the cube of wind speed (Fischer et al. (1979)), we use the cube of wind speed as our signal. To take advantage of the hourly resolution of the wind product, we use the daily average of cubed hourly wind speed (Fig. 2a).

Vertical eddy diffusivity

The vertical eddy diffusivity (VED) represents the strength of mixing in the system (Soontiens and Allen (2017)) and depends on the choice of vertical turbulence closure scheme. SalishSeaCast uses a k-ϵ configuration of a generic length scale turbulence model to estimate sub-gridscale turbulent processes (Umlauf and Burchard (2003)), with background vertical eddy viscosity and diffusivity both set to 10^{-6} m^2 s^{-1}. We report a daily depth-averaged value (Fig. 2b). Though average vertical eddy diffusivity reflects all sources of mixing and stratification present in the system, it is dominated by barotropic tidal activity, and we expect it to be highest at tidal mixing hotspots (Crean (1978)).

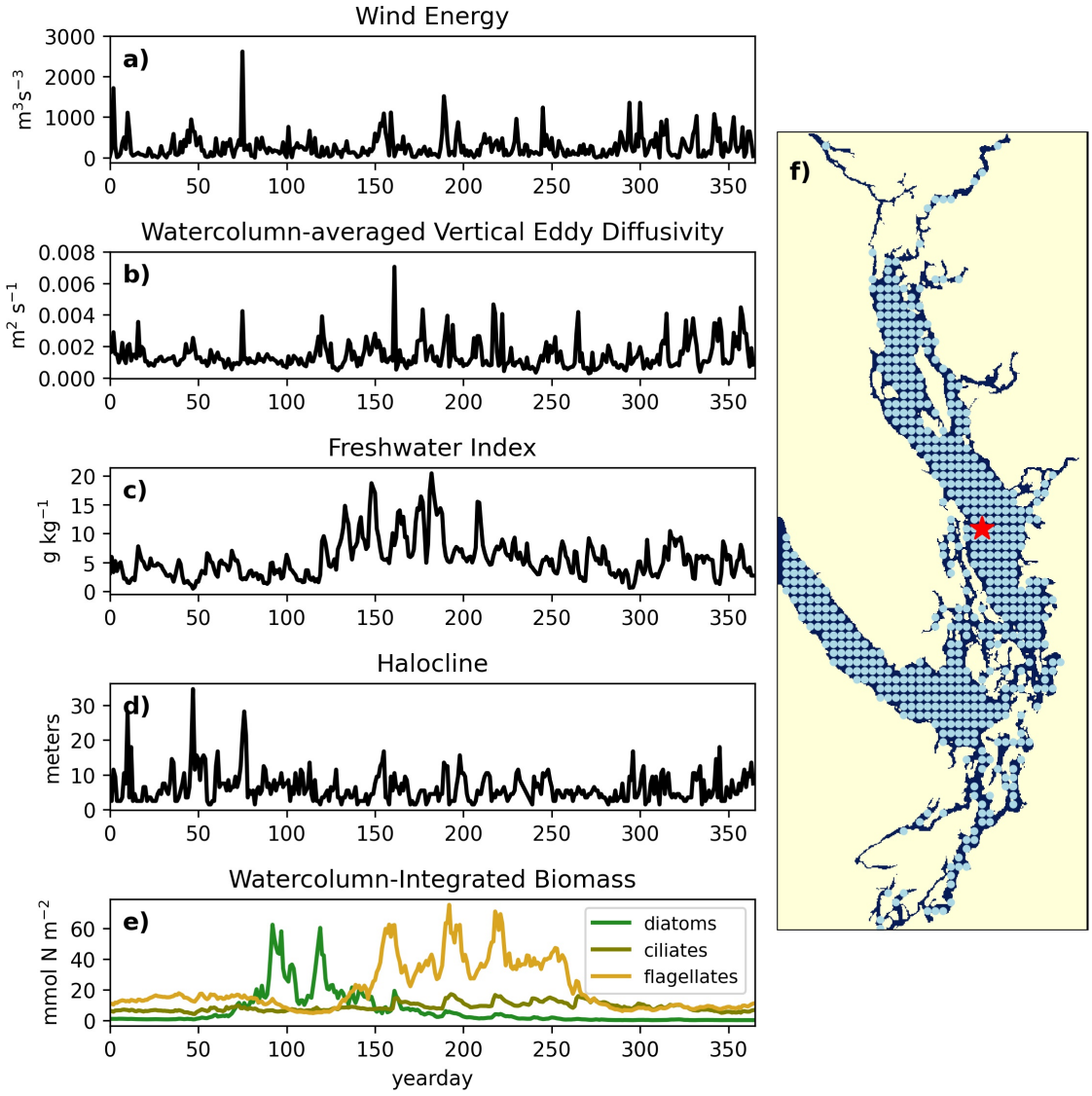

Figure 2. Example yearly signals of clustered physical and biological factors from one station in the CSoG (red star). The physical signals are as follows: a) wind energy (m^3s^{-3}), b) watercolumn-averaged vertical eddy diffusivity ($m^2\ s^{-1}$), c) freshwater index (g kg^{-1}), d) halocline (m). The biological signal is watercolumn-integrated phytoplankton biomass (mmol N m^{-2}), separated by functional group (diatoms, ciliates, and flagellates). The remaining 570 stations used in the clustering are shown as blue points. Depth-integrated phytoplankton biomass signals are combined in series for clustering (see Figure 8).

Freshwater index

The freshwater index (Fig. 2c) is intended as a proxy for freshwater influence on the watercolumn at a given station, and is expressed as the salinity difference between the mean of the surface 4 meters of the watercolumn and the salinity at depth 50 m,

in units of g kg^{-1}. This metric may be thought of as a salinity stratification metric. Where the watercolumn is shallower than 50 meters, the salinity at 50 m at the nearest model point that is 50 m deep is used. Similar metrics have been used as indicators of stratification in the region (Suchy et al. (2019); Masson and Peña (2009)), but were typically based on the difference in water density between the surface and the deep waters; here we isolate the impact of salinity alone by using a salinity-based metric. As salinity dominates stratification in this region (LeBlond (1983)), we expect clusters derived from a salinity-based clustering to be broadly similar to those derived from a density-based clustering. The value of 50 m was chosen because the majority of the Salish Sea is more than 50 meters deep; however, we do not expect the results to change dramatically if a different depth were to be chosen.

Halocline depth

The halocline depth (Fig. 2d) is defined as the depth of the maximum salinity gradient in the water column, which is estimated by finding the salinity difference of two adjacent cells in the vertical dimension and reporting the depth at the midpoint of the two cells that have maximum salinity gradient in the watercolumn at a given station.

Phytoplankton biomass

We extract daily-average depth-integrated phytoplankton biomass (mmol N m^{-2}) for each of the three phytoplankton functional groups to form three signals (i.e. one year-long daily-resolution timeseries of depth-integrated phytoplankton biomass for each of the three phytoplankton functional groups, (Fig. 2e)). These signals are then connected in series to form an overall phytoplankton biomass signal that differentiates by functional group - thus, functional group identity, not just total phytoplankton biomass, is a factor in our clustering. Furthermore, our chosen metric of functional-group-differentiated phytoplankton biomass will capture functional-group specific responses to different habitat characteristics.

2.3 Clustering method

We use Ward's method (Ward Jr (1963)), a type of hierarchical clustering method, to cluster our data. Broadly, clustering methods are a subset of unsupervised machine learning methods used to reveal the underlying structure of a dataset by grouping similar data points. In hierarchical clustering methods, every datapoint is initially a single-point data cluster. At each step of the clustering, the two 'closest' clusters are merged into a new cluster; this process is repeated until all points have been merged into a single cluster. Metrics of closeness vary between hierarchical clustering methods - while some methods use variations of the definition of the physical distance between clusters as a clustering criterion, Ward's method analyzes changing intracluster variance, or the "loss of information" (Wishart (1969)) if they were to merge into a single cluster. In Ward's method, at each step, the clusters whose merging results in the lowest increase in intracluster variance are combined.

Many hierarchical clustering methods exist; of these we chose Ward's method because the algorithm is straightforward to implement and compares favourably to other hierarchical clustering methods with regards to performance in identifying structure in known clusterings (e.g. Mangiameli et al. (1996)). We perform hierarchical clustering using Ward's method on

each of the five signals independently. For each signal, the clustering is done four times (once for each of the four years 2013-2016), and the results for the four years are then compared to assess interannual variability in the patterns found.

Cluster number selection

A common challenge in the application of clustering methods is the selection of cluster number, as the clustering algorithm can produce anywhere between 2 and N clusters (where N is the number of signals being clustered). Typical approaches include choosing a cluster number where the difference in the mean signals of the found clusters when going from cluster number N to cluster number N+1 is maximized. In our case, attempts to use objective metrics to determine cluster number, such as the Davies-Bouldin, Silhouette, or Calinski-Harabasz criteria (Maulik and Bandyopadhyay (2002)) typically identified only two clusters in a given dataset (not shown). Though these may be the most prominent clusters, meaningful structure in the data persists at larger cluster numbers. Ultimately, our approach was to visualize several possible clustering outputs, with cluster number N varying from 2 to 15, and to visually compare how the spatial structure of the patterns changed with increasing cluster number (e.g. Fig. C1). In all variables, the same typical structure emerged at a relatively low cluster number (eg, N = 3-5) and persisted with increasing cluster number in all years. To facilitate comparison of clusters between years, we chose an N=5 for all years for all clusters, and are confident that the structures described are robust to a selection of a variety of cluster numbers.

Interannual cluster persistence

Visually, it is immediately apparent that similar spatial structure in the clusterings of a single variable persists interannually. To formalize the interannual persistence of a single cluster between years, as well as spatial commonality of different variables, we establish a simple nondimensional cluster commonality metric (CC). For two clusters A and B, the cluster commonality CC_{AB} is defined as:

$$CC_{AB} = \frac{|A \cap B|}{0.5(|A| + |B|)}$$

For any two clusters, CC varies from 0 (clusters of any size with no stations in common) to 1 (two clusters of equal size with all stations in common) and may be used to compare clusters of unequal sizes. We use this metric to compare the persistence of clusters of individual variables between years, as well as the cluster persistence between different variables in a given year (Fig. C2).

3 Results

We describe the main physical-oceanographic subregions in the domain (CSog, NSoG, and JdF) determined by clustering the physical factors and interpret our results in the context of previous work. We also consider some tidal mixing hotspots highlighted in the derived map of vertical eddy diffusivity. Our results here typically extract the main known general physical-oceanographic features of this coastal sea. We then describe the observed spatial regions in biomass, which are remarkably

cohesive, in the context of these physical factors. In the discussion, we propose some mechanisms through which the physical factors likely shape the biological structures seen here.

3.1 Central Strait of Georgia

The physical-oceanographic dynamics of the CSoG are dominated by stratification due to Fraser river runoff, which is easily visible in the derived clustering of the freshwater index (Fig. 3). Spatially, in all four years of our analysis, the highest freshwater index is seen near the mouth of the Fraser River and in Howe Sound (cluster 1/gold), and then radially decreases in bands (cluster 2/grey, cluster 3/sky blue) outward from this maximum. The tendency of the surface Fraser River plume to move north from the mouth of the Fraser due to the Coriolis force (Liu (2014)) is also easily observable in this visualisation. The stratifying tendency of the Fraser river (and of other major rivers) is then reflected in the clustering of the halocline signals (Fig. 4). The CSoG (cluster 3/sky blue) has consistently shallow haloclines with only limited seasonal variability ($\sim$5m in summer to $\sim$7m in winter). These shallow, stable haloclines also persist in most of the Puget Sound, owing to the influence of the Skagit River, and in the northern fjords with large rivers at their head (Toba Inlet, Bute Inlet, and Howe Sound), and the influence of these rivers is reflected in the clustering of the freshwater index. Because rivers other than the Fraser are forced by climatology in the model, the potential effects of the interannual variability of their hydrographs are not seen here.

In the wind clustering, the boundary between the CSoG and NSoG is farther south than that seen in the clusterings of freshwater index and halocline depth (Fig 5). Though winds in all clusters are highly episodic, all wind clusters show a marked decrease in wind energy during the summer months (Fig. 6) - this change in mean signal magnitude and variability is most pronounced in the NSoG (cluster 4/red), which consistently shows $\sim$2 times higher wind energies in the winter months than in the summer months. In contrast, the CSoG (cluster 3/sky blue) shows lowest variability between summer and winter energy magnitudes. Summer wind energies are actually higher in the CSoG than in the NSoG, likely due to the long wind fetch length in the CSoG, as summer winds in the Salish Sea are predominantly northerly (Thomson (1981); Moore-Maley and Allen (2022)). Average vertical eddy diffusivity is lowest in the CSoG (Fig. 7), owing likely both to high stratification and to comparatively low tidal currents (Thomson (1981)), consistent with the historical idea of the Salish Sea as a system of relatively quiet basins interconnected by dynamic sills (Ebbesmeyer and Barnes (1980)).

3.2 Northern Strait of Georgia

As expected, the influence of the Fraser river is lower in the NSoG as the region is farther away from the rivermouth (Fig. 3). The resulting lower stratification is reflected in deeper and more variable haloclines in all seasons (on average, $\sim$ 10m in summer to $\sim$20m in winter) (Fig. 4). A striking feature in the clustering of the freshwater index signal and halocline signals in the NSoG and the CSoG is the dissimilarity of year 2016 to other years, reflected in a lower cluster persistence metric in this year (Fig. C2). Maximum Fraser River discharge (freshet) during 2016 was remarkably low, in the lowest quartile of discharge on record, reaching only $\sim$8,000 m^3s^{-1}, or roughly 2/3 of the magnitude of the 2013-2014 freshets, which were both in the highest quartile (Fig. C3). Interestingly, the mean freshwater index signal for each cluster in 2016 remains similar to the means for other years, as does the spatial extent of the most river-influenced cluster (cluster 1/gold), but the medium

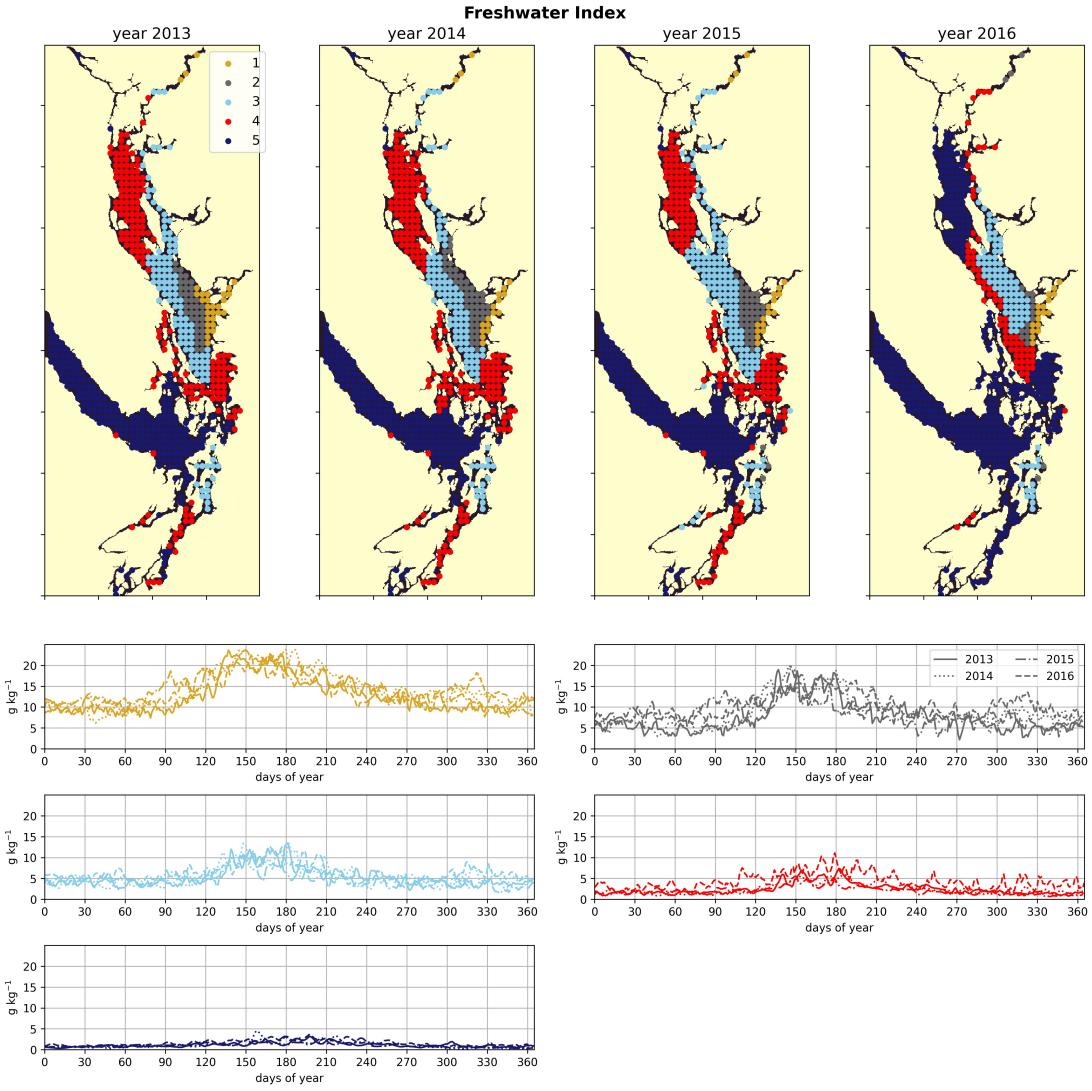

Figure 3. Clustering of the freshwater index signal (Section 2.2). As expected, areas near the mouth of the Fraser river have the highest freshwater index, with the freshwater plume turning north due to the Coriolis force, and the index decreases in bands from this maximum. Elevated freshwater index can also be seen in the vicinity of the Skagit river in Puget Sound and at the head of Toba Inlet, Bute Inlet, and Howe Sound, which contain glacial rivers. The magnitude of the freshwater index in the different clusters does not vary significantly interannually, but the spatial extent is diminished in year 2016, which had lowest freshet magnitude of the four years. In all clusters, the freshwater index peaks at the same time as the Fraser freshet does for a given year.

freshwater-influenced clusters (cluster 2/grey, cluster 3/sky blue) extend less far from the river mouth. As a result, the NSoG clusters with JdF in this year.

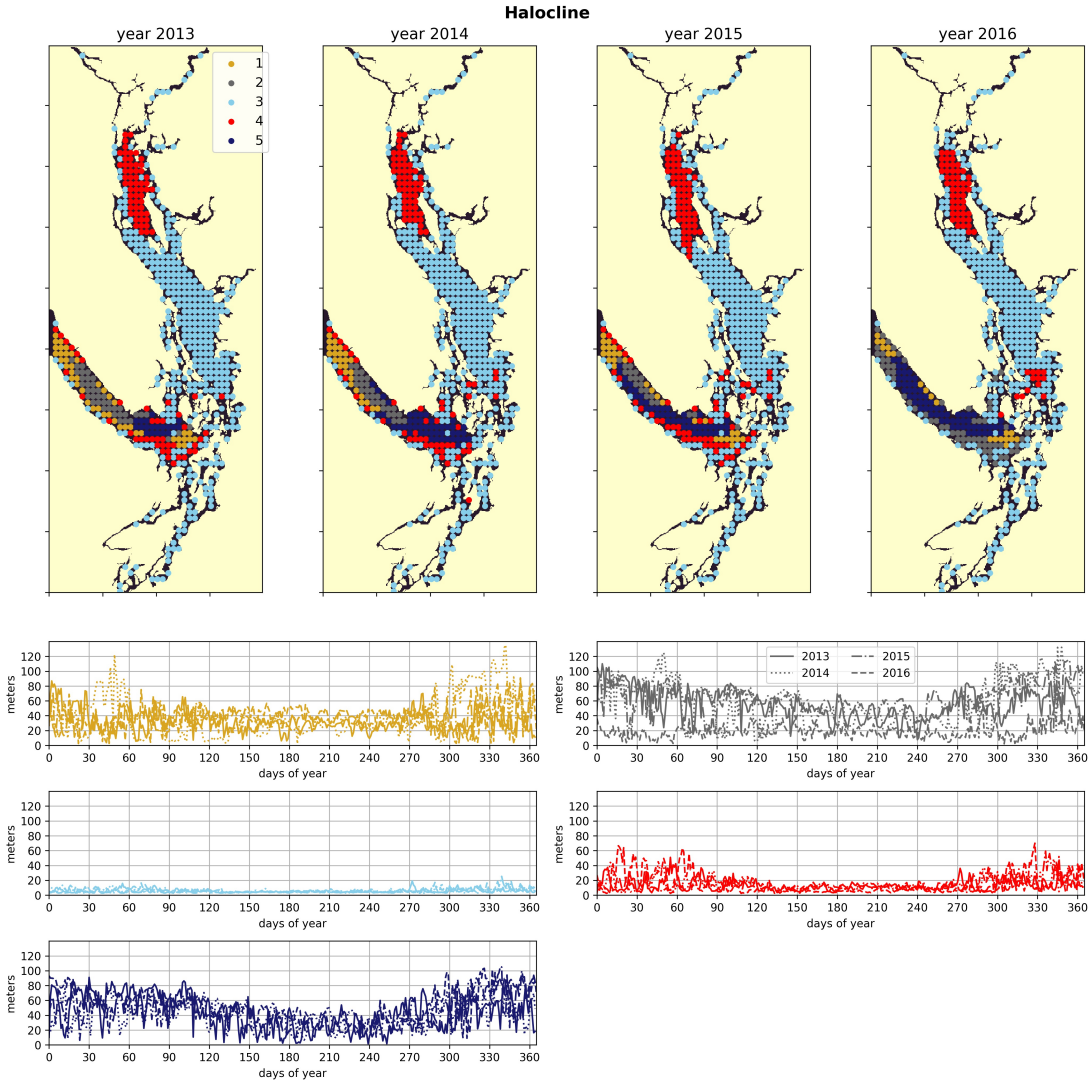

Figure 4. Clustering of the halocline signal, defined as the depth of the maximum salinity gradient. The largest region (cluster 3, light blue) is the freshwater influenced CSoG, with shallow (<10 m) haloclines and limited variability between seasons. Similar halocline dynamics are seen in Puget Sound and at the head of Toba Inlet, Bute Inlet, and Howe Sound, which contain glacial rivers. Significantly deeper and more variable haloclines are found in the NSoG (cluster 4, red), commonly deeper than 40 m in winter. The deepest and most spatially variable haloclines occur in the center of the JdF (clusters 1, 2, and 5), with nearshore regions of the JdF clustering with the NSoG in most years (cluster 4).

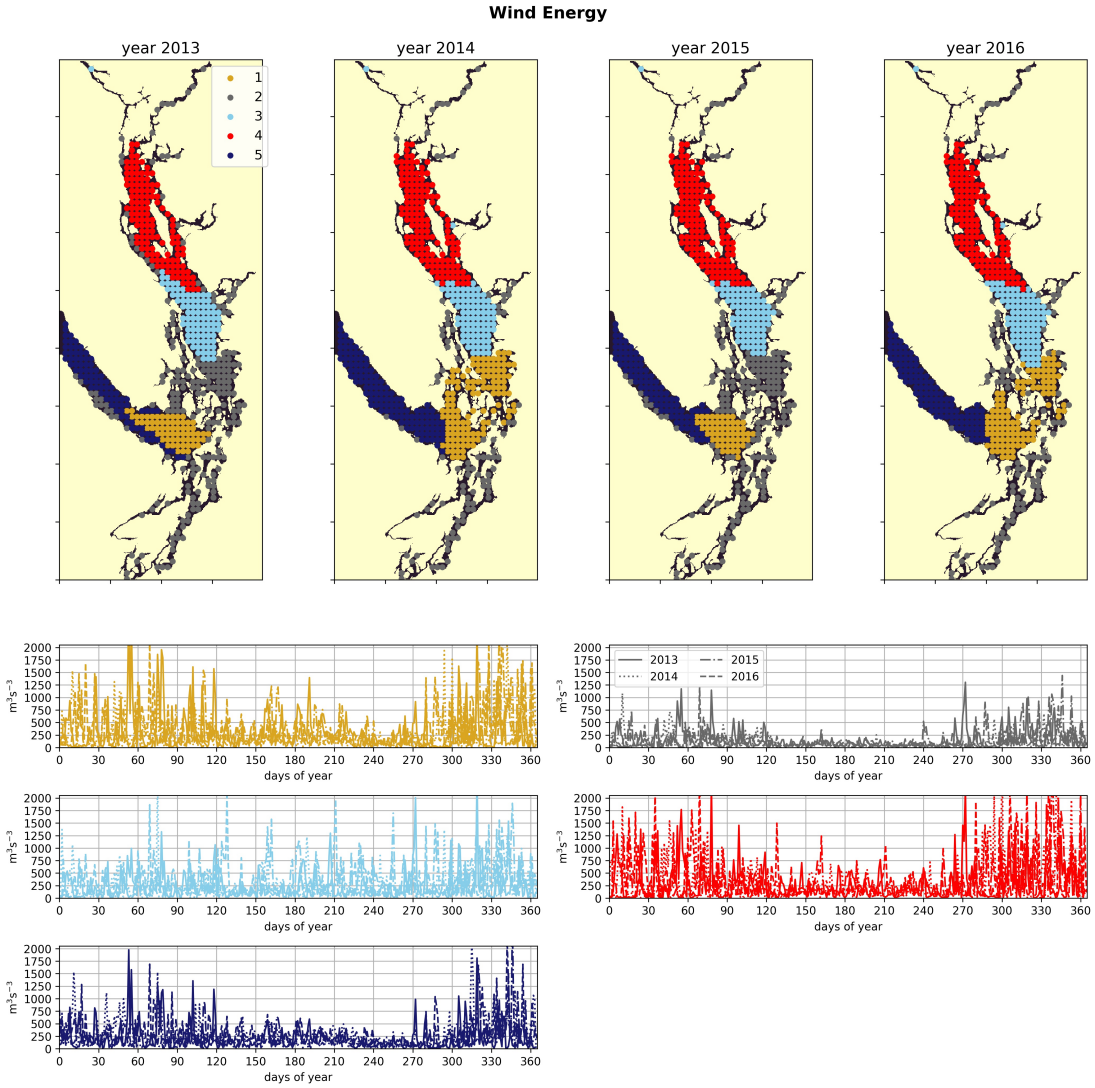

Figure 5. Clustering of the daily-average wind energy signal. Though spatial cluster boundaries are consistent, wind energy in all clusters is highly episodic, and all wind clusters show a marked decrease in wind energy during the summer months. Nearshore ares have lowest wind energy, owing to low fetch. Summer wind energies are higher in the CSoG than in the NSoG.

3.3 Juan de Fuca Strait

Dynamics in Juan de Fuca Strait are broadly characterized by limited local freshwater influence, though a small increase in freshwater index is visible in the summer months (Fig. 3, 6), in part because of the surface advection of freshet-driven water from the CSoG due to the estuarine circulation (Thomson et al. (2007)). The limited freshwater stratification, accompanied by a large tidal range, results in deep and variable haloclines (Fig. 4). The larger tidal velocities here are also reflected in slightly

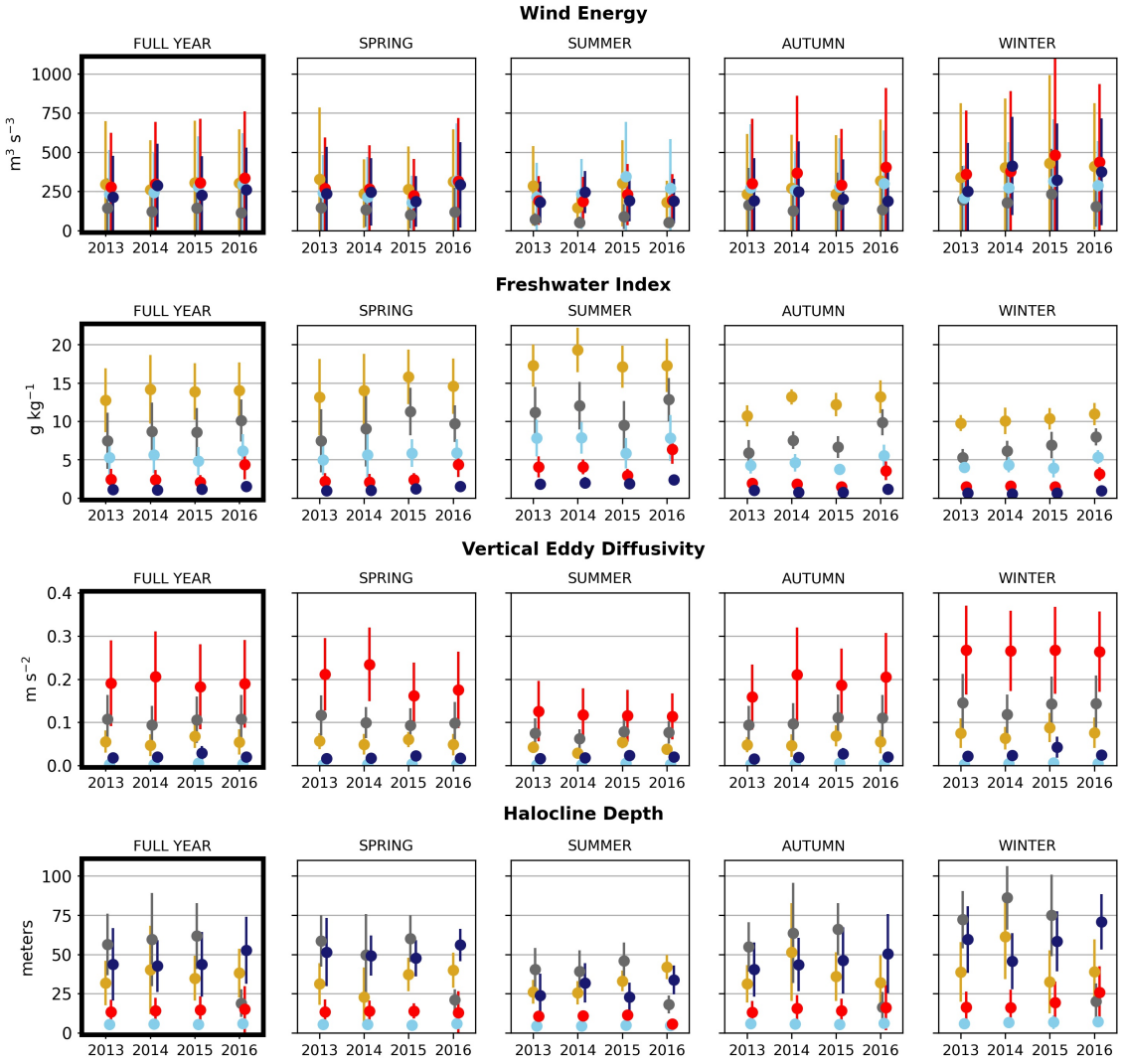

Figure 6. Seasonal means of the physical signals. Seasons are defined as follows: Winter is Dec-Feb, Spring is Mar-May, Summer is Jun-Aug, Fall is Sep-Nov. The temporal standard deviation of the seasonal mean signal for each cluster is shown.

255 higher watercolumn-averaged VED (Fig. 7). Interestingly, in 2015, the VED in much of JdF clusters with the SoG, possibly due to the inhibition of water column mixing by higher thermal stratification of the system due to the significant marine heatwave in the North Pacific in the years 2013-2015 (Gentemann et al. (2017)), whose effects were most pronounced in the Salish Sea in 2015 (Chandler et al. (2016)). The dissimilarity of year 2015 to other years is reflected in the cluster persistence metric (Fig. C2).

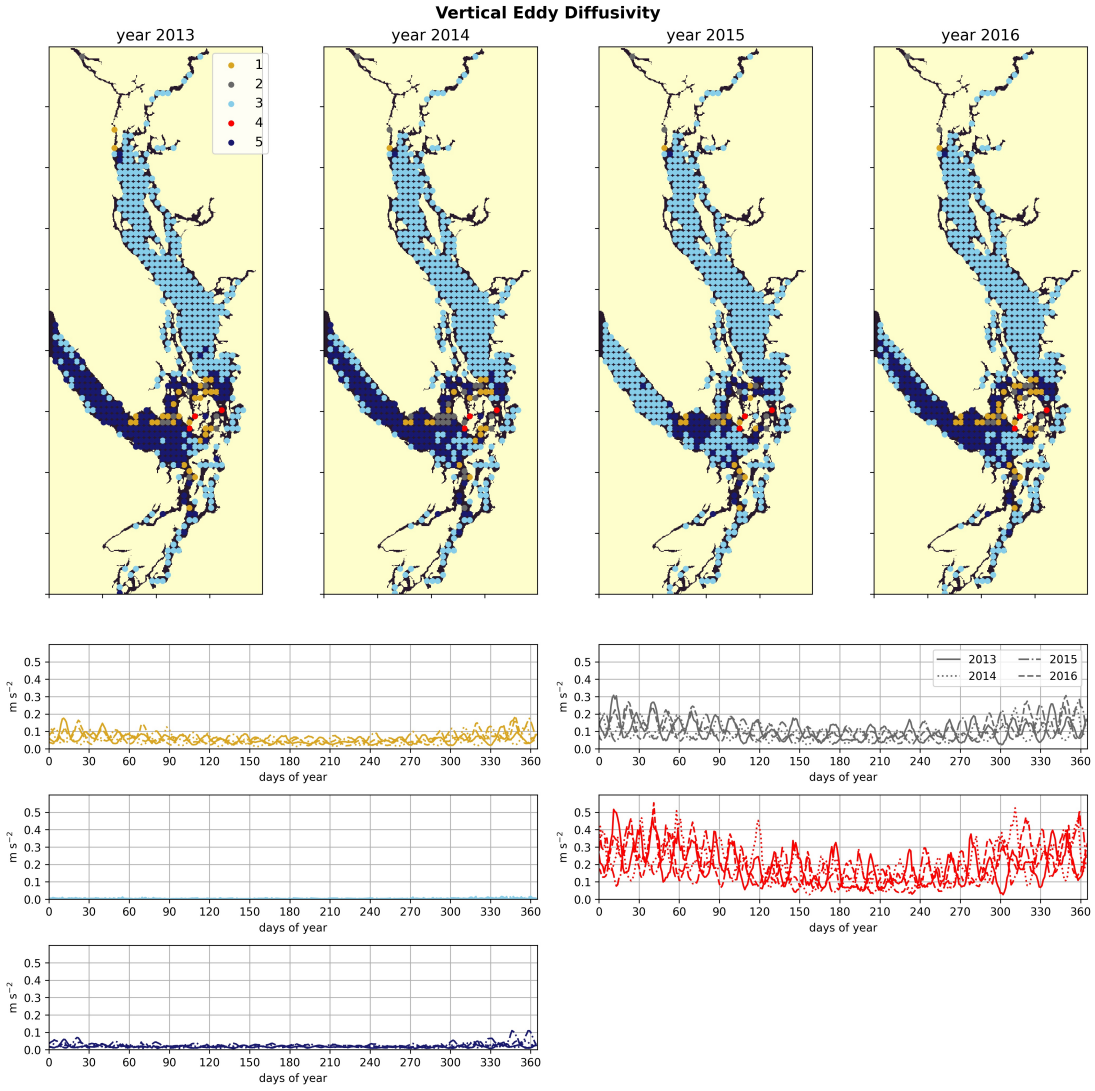

Figure 7. Clustering of the daily depth-averaged vertical eddy diffusivity signal. The domain is split into two major regions: the Strait of Georgia, which has universally low vertical eddy diffusivity, and Juan de Fuca Strait, with comparatively slighly higher VED due to stronger tidal currents. VED hotspots of various magnitudes are consistently found at tidal mixing hotspots, including Discovery Passage near Seymour Narrows and Haro Strait near the San Juan islands.

3.4 Tidal Mixing Hostpots

Watercolumn-averaged vertical eddy diffusivity in the Salish Sea is dominated by tidal mixing activity (Crean (1978)), allowing clustering VED to uncover dominant tidal hotspots. VED varies by three orders of magnitude in the model domain (Fig. 7). As expected, this metric reaches its maximum in the Haro Strait region, as well as in parts of Puget Sound, for example in

Admiralty Inlet, near known tidal mixing hotspots (Ebbesmeyer and Barnes (1980); LeBlond (1983); Moore et al. (2008); Deppe et al. (2018)). Two stations in northern Johnstone Strait and the Discovery Passage region also exhibit heightened VED in all four years, consistent with the high observed tidal velocities near Seymour Narrows in this region (Thomson (1981)). Fourier analysis of the annual vertical eddy diffusivity signals also shows local maxima in energy at weekly and fortnightly frequency (not shown) in all four years in all five clusters, consistent with the role of tides as the dominant source of mixing energy in the system (Crean (1978)).

The by-cluster seasonally-averaged means and standard deviations of average VED are consistent interannually (Fig. 6, 7). The same three stations in the San Juan islands (cluster 4/red) report highest VED in all four analyzed years, exhibiting maxima that are almost a factor of two larger than the next largest signal (cluster 2/grey). In the highly-variable Haro Strait and Johnstone Strait regions, the spatial frequency of our sampling likely plays a role in our derived map of tidal mixing hotspots - as we sample only approximately every 100th horizontal model coordinate, we likely miss other high-VED model points in this subregion, especially channels that have width-scales comparable to our model resolution (0.5km), for example the intricate channel passages of the San Juan and Discovery Island groups in the Haro and Johnstone Strait regions, which are known tidal mixing hotspots (Fig. 1). Analysis of tidal mixing hotspots is not the focus of this work, but a full characterization of this tidally-mixed zone using a more refined clustering approach may be an interesting focus of future work.

3.5 Biomass of primary producers

A similar biological clustering arises in all four years (Fig. 8, Fig. C2). The boundaries of this clustering coincide broadly with the three major oceanographic subregions discussed above. The largest cluster (the CSoG - cluster 3/sky blue) is characterized by diatoms blooming first, followed by a transition to flagellate abundance in the summer months. In all four years, a functionally distinct NSoG region (cluster 4/red) arises, with sharp, epsisodic spikes in summer diatom biomass and diminished flagellate biomass. JdF (cluster 5/dark blue) reaches maximum biomass later in the year and, like the NSoG, shows a persistence of summer diatoms and diminished summer flagellate biomass. In contrast to the NSoG, where diatom biomass diminishes between episodic spikes, diatom biomass in JdF typically remains above 20 mmol N m^{-2} throughout the spring and summer seasons, with occasional spikes to higher biomass.

These three main regions have roughly similar mean seasonal biomass, with interannual variability larger than variability between clusters; the main differences between them are in the relative abundances of different functional groups and in the temporal characteristics of the phytoplankton biomass. Nearshore areas cluster together (cluster 2/grey) and have low depth-integrated biomass because they are limited by shallow depth. The largest depth-integrated biomass in the model in both the diatom and flagellate groups is found in the tidal mixing region of Haro Strait (cluster 1/gold).

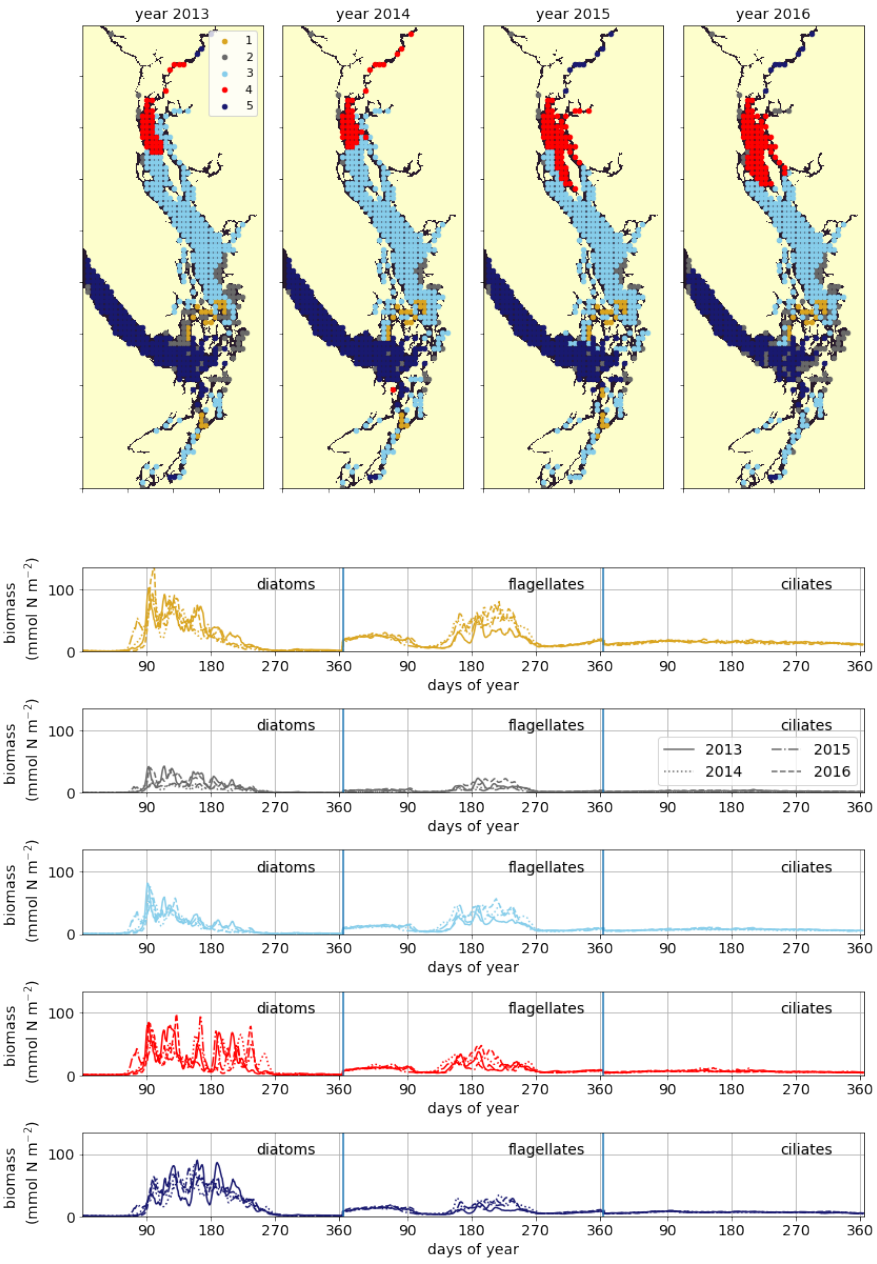

Figure 8. Clustering of vertically integrated phytoplankton biomass separated by model-defined functional group (diatoms, followed by flagellates, then ciliates). The domain is split into the CSoG, NSoG, and JdF, each of which exhibit distinct phytoplankton dynamics (see Section 3.5 and Discussion).

4 Discussion

We now consider the regional phytoplankton structure in the context of previous observational and modelling studies and discuss some mechanisms underlying the observed patterns. We focus on the three main regions found by the biological clustering (the CSoG, the NSoG, and JdF).

4.1 The North vs. the Central Strait of Georgia

In the model, the NSoG shows only slightly higher depth-averaged phytoplankton biomass in all seasons than the CSoG (Fig. 9). This biomass is consistent with the in-situ study of Masson and Peña conducted between 2001-2007 in this region, which shows lower surface chlorophyll but a deeper phytoplankton growing zone in the NSoG leading to slightly higher depth-integrated chlorophyll concentrations in the northern region in all four seasons of sampling (Masson and Peña 2009, henceforth MP09, Table 2). Remote sensing observations also show significantly lower surface chlorophyll in the NSoG throughout the year, as well as finding anomalously high surface chlorophyll concentrations in 2015 that are not reproduced by the model (Suchy et al. (2019)). The majority of the modelled biomass difference between the NSoG and CSoG occurs in the subsurface maximum around 6-8 meters in depth (Fig. C4), where it cannot be detected by remote sensing. Previous modelling (of years 2007-2009) found somewhat higher depth-integrated phytoplankton biomass in the CSoG throughout the year, but with significant spatiotemporal variability (Peña et al. (2016)). However, a recent year-round in-situ campaign in several parts the Strait of Georgia found no meaningful difference in depth-integrated chlorophyll between the NSoG and CSoG (Pawlowicz et al. (2020)).

Together, these studies suggest that the difference between the two regions with respect to total depth-integrated biomass is subtle. However, we find a substantial difference in the modelled phytoplankton functional group composition and the temporal scale of variability of the phytoplankton signal between regions. In both regions of the Strait, the opportunist-class diatoms, which have the highest growth rate and highest nutrient requirements, peak first (typically in late March, though with considerable variability (Allen and Wolfe (2013))) and form the majority of the phytoplankton biomass in the spring (fig. 8, 9). In the CSoG, the model then transitions to higher biomass of gleaner-type flagellates around day 150, near the beginning of June, and flagellates continue to exhibit high summer biomass in this region (Fig. 8, 9). In contrast, the NSoG continues to exhibit episodic short-lived peaks of high opportunist-type phytoplankton biomass, represented by diatoms, throughout the summer so that in all years except 2016, diatoms make up the majority ($\sim$55-60$\%$) of summer phytoplankton biomass in this region.

In the Strait of Georgia, significant evidence of high summer biomass near strong mixing zones or in response to mixing driven nutrient delivery exists. For example, early surveys of the system find high chlorophyll associated with dynamic frontal regions in the northern and southern ends of the SoG (Parsons et al. (1981)), and consequently warn against drawing firm conclusions about the nature of phytoplankton abundance and variability from episodic sampling in shifting frontal zones. Nutrient delivery via episodic tidal mixing events near Discovery Passage has been linked to increased biomass (e.g. Parsons et al. (1981); Haigh and Taylor (1991)) and modelled primary productivity (Olson et al. (2020)). Though this phenomenon has

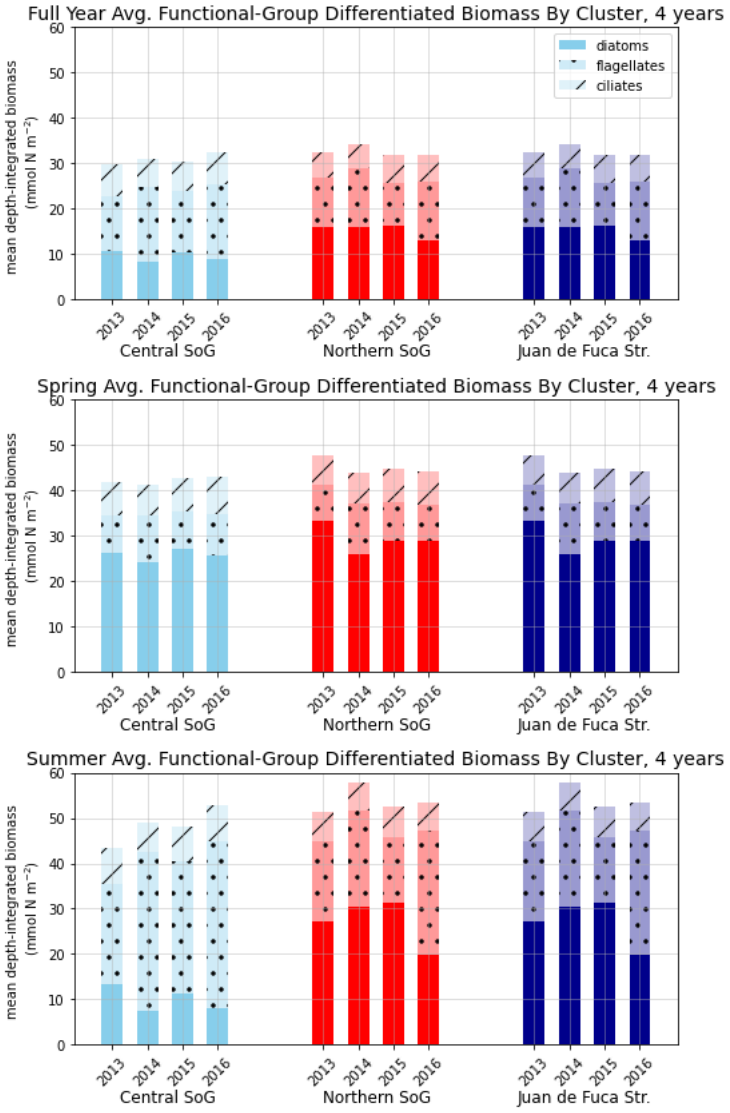

Figure 9. Depth-integrated phytoplankton biomass for the three main biological clusters (CSoG, NSoG, and JdF), differentiated by functional group. The annual, spring, and summer means of the derived clusters are shown for all four modelled years. All three clustered regions have similar total biomass, which stays relatively consistent interannually, but functional group composition varies by cluster, with higher summer diatom abundance in the NSoG and JdF than in the CSoG. Spring is defined as March-May, and Summer is June-August.

been recorded in the CSoG as well (Yin et al. (1997); St. John et al. (1993)), higher stratification may dampen the magnitude of the nutrient pulses. Sudden introduction of abundant nutrients is expected to favour the opportunist functional group represented by diatoms over the slower-growing gleaner functional group represented by flagellates, as is seen in our clustering (Cloern and Dufford (2005); Dutkiewicz et al. (2009)). A recent four-year timeseries of phytoplankton composition data at a station

near Quadra Island in the NSoG supports this idea by showing episodic blooms of summer diatoms after wind events (Del Bel Belluz et al. (2021)). Indirect evidence of episodic high biomass, sometimes following wind events, has been observed elsewhere in the NSoG (Evans et al. (2019); Mahara et al. (2021)).

We suggest that our results reflect a controlling influence of stratification on phytoplankton biomass and community structure. Strong stratification concentrates phytoplankton biomass in a thin well-lit surface layer while limiting supply of nutrients after the initial biological drawdown. In the model, these conditions favour high abundance of the gleaner-flagellate group. In the NSoG, nutrient drawdown also occurs, but episodic wind events lead to stronger upwelling and mixing due to the comparatively weaker stratification and inject sharp pulses of nutrients into the near-surface, leading to sharp, short-lived diatom blooms (Moore-Maley and Allen (2022)). In contrast, in the CSoG, despite stronger summer winds, strong stratification continues to favour gleaner-type organisms. Faster-growing opportunist-diatoms tend to outcompete gleaner-flagellates when sufficient nutrients and light are available, but inherent variability in the physical environment promotes coexistence (Anderies and Beisner (2000)). The result is only a modest, if any, change in biomass but a significant change in functional group composition and temporal variability between the NSoG and CSoG.

Peña et al. find higher biomass in the CSoG due to the deeper nutricline in the NSoG (Peña et al. (2016)). We find instead that the increased mixing in the NSoG provides increased nutrients and that biomass in both regions is about the same. These two views are not directly reconcilable and which view is more representative of actual conditions depends on accurately capturing the balance of between the action of mixing as a source of nutrients and mixing as a source of light reduction and phytoplankton dilution.

4.2 Comparison with in-situ phytoplankton functional group observations

The extent of phytoplankton diversity in the Salish Sea cannot be strictly condensed into the three functional groups represented in this model. The divide of the phytoplankton functional groups in the model does not precisely correspond to a split between diatoms and all types of flagellates. For instance, silicoflagellates (class Dictyochophyceae) might align with the diatom class based on silicon utilization. For this reason, we discuss these classes in terms of competition between opportunist-type primary producers with high nutrient needs and high light needs and gleaner-type primary producers with capacity to persist at lower nutrient and light levels. For this study, the requirement is capturing the overall regional biomass patterns and function. Here we provide a brief comparison between our results and available in-situ phytoplankton functional group observations.

In-situ measurements of the relative abundance of phytoplankton functional groups remain rare in the Salish Sea, and tend to be sparse in both space and/or time. In situ observations represent a snapshot at a single station and depth, while model output instead presents the average of a much larger volume (the discrete model cell). A recent well-temporally-resolved four-year (2015–2018) time series of phytoplankton biomass and composition, derived from high-performance liquid chromatography (HPLC) analysis of phytoplankton pigments, taken at a single station in the NSoG (Del Bel Belluz et al. (2021)), provides a starting point for such a comparison. The in-situ data, taken from a depth of 5 meters, show diatom-dominated blooms with varying start dates in the spring season, followed by a transition to a regime where flagellate-type groups (chiefly prasinophytes and cryptophytes) make up the majority of phytoplankton biomass, but diatoms remain present (Del Bel Belluz et al. (2021),

Fig. 4, Fig. 5). Episodic later-summer diatom blooms occur in three of the four observational years, corroborating the modelled later-summer NSoG diatom blooms seen in this study.

Local phytoplankton composition data derived from shipboard observations from spring, summer and autumn cruises spanning the Juan de Fuca Strait and both the CSoG and the NSoG are also available as a technical report by Nemcek et al. (2020). Unlike the Del Bel Belluz study, these data are relatively well-resolved in space but less resolved temporally; in a given year, typically only one day of observations is available for each station. The relative abundance of phytoplankton functional groups in the three regions is thus interannually variable (Nemcek et al. (2020), Fig. 39-2), and in contrast to the Del Bel Belluz study, the data show only limited summer presence of diatoms in the NSoG in either of the years overlapping with our study (2015 and 2016). These observations contradict trends seen in our model. Simultaneously, these data show summer diatom dominance in the well-mixed Haro Strait region, corresponding to our tidal mixing region, which echoes trends we see in this study. Taken together, these two in-situ phytoplankton composition studies each provide some corroboration of patterns we see in the model, but differ from each other on summertime diatom representation in the NSoG. Combined with the modelling, these three perspectives on phytoplankton composition and biogeography each represent different spatial and temporal scales. The questions raised by their contrasting findings highlight the need for both modelling and observational work to provide a holistic view of the local biophysical dynamics.

Our modelling study is necessarily subject to limitations. For example, very high biomass shown in the tidal mixing region (Fig. 8, cluster 1 -gold) could be an artifact of slower phytoplankton mortality rates, at least at times, than occur in nature, with phytoplankton mixed deep into the water column and persisting too long. Such a rate imbalance would affect the response to mixing described above. Available observations support model phytoplankton levels in this region but are limited to the upper water column. Because model-data agreement in biomass and nitrate is strong in these regions, we believe the mechanism of nutrient delivery by wind events in the less stratified north leading to dominance of faster-growing phytoplankton is robust.

4.3 Juan de Fuca Strait

Our results suggest that the mean annual average depth-integrated biomass is about the same in all three physical regions, including the well-mixed, weakly-stratified JdF. In contrast, previous studies suggested a lower biomass in JdF (Masson and Peña (2009); Peña et al. (2016)) due to a deeper nutricline. However, recent in-situ chlorophyll and nutrient data (2013-2016) support our result. In fact, the evaluation suggests that, at dates and locations where observations are available, the model slightly underestimates observed biomass in Juan de Fuca Strait (Fig. A2, Table A2).

One factor contributing to the difference between these conclusions may be the vertical structure in the biomass observed by both MP09 and the model. In MP09, the spring phytoplankton biomass is much more prominent in the CSoG and NSoG than in JdF. The spring biomass exhibits a strong subsurface maximum ($\sim$10 meters in the chlorophyll observations) and persists relatively deep into the watercolumn (up to 40 meters). However, though overall concentrations reported in MP09 are lower in all seasons in JdF, observed chlorophyll concentrations $>= 1$ mg m^{-3} persist at deeper depths in most seasons in Juan de Fuca than in both regions of the Strait of Georgia (up to 50 meters in the spring, summer, and fall), and in summer and fall, the NSoG exhibits slightly deeper phytoplankton persistence compared to the CSoG. We replicate these trends in general vertical

structure (Fig. C4), with a prominent subsurface maximum at ∼6-8 meters and phytoplankton biomass mixed deeper in Juan de Fuca Strait than in either region of the Strait of Georgia.

This vertical structure likely leads to a dilution effect - even when phytoplankton concentration at a given depth may be lower in Juan de Fuca Strait than in the Strait of Georgia, overall depth-integrated biomass may be simultaneously higher. This deep biomass is less likely to be captured by sampling campaigns, potentially leading to an underestimation of the phytoplankton biomass of the region as a whole. Furthermore, because of the interannual variability in spring bloom timing and differences in spring bloom timing between the Strait of Georgia and Juan de Fuca Strait (discussed further below), the spring in-situ survey that captures high biomass in Strait of Georgia may be too early to observe the full extent of the spring bloom in Juan de Fuca Strait.

In the NSoG and CSoG, the derived biological signals suggest a regime where stable growing conditions in the spring transition to varying degrees of summer nutrient limitation which are interrupted by episodic nutrient delivery, more frequently in the NSoG. In contrast, the diatom growth curve in JdF suggests a light-limited environment year round, consistent with the established understanding (Mackas and Harrison (1997)). Nutrients are rarely limiting in JdF, owing to the plentiful supply of oceanic nitrate (Sutton et al. (2013)) and stronger watercolumn mixing in this region demonstrated in the VED clustering (Fig. 7). One factor that may potentially enhance growing conditions in the summer season here is the advection of a freshwater lens from the Strait of Georgia via the surface estuarine circulation (Pawlowicz et al. (2007); MacCready et al. (2021)). This advection is visible as a slight increase in the summer freshwater index in JdF (Fig. 6) and may contribute to increased watercolumn stability, and hence light availability and favourable growing conditions, in this period.

4.3.1 Spring Bloom Timing

The timing of the first substantial increase in phytoplankton biomass (the spring bloom), in the Salish Sea varies considerably inter-annually and is driven by different factors, primarily wind speed and cloud cover, and secondarily temperature and freshwater discharge (Allen and Wolfe (2013)). While we do not evaluate spring bloom timing here, considering the spatial variability of the onset of the spring bloom throughout the domain may deliver insights regarding the functioning of the different regions. For the purposes of this informal exploration, at each station we define the spring bloom as the first peak in depth-integrated diatom biomass that is at least 30% of the maximum annual diatom biomass at that station. Earlier spring bloom initiation in the CSoG with respect to the NSoG was seen in multiple years of satellite observations (Suchy et al. (2019)). In our results this progression within the SoG is almost indistinguishable and is followed by later blooming in the JdF. The late bloom timing in JdF was likely driven by stronger mixing limiting light availability later into the year in JdF region (Fig. 10), consistent with the functional differences between JdF and the NSoG and CSoG discussed above.

This preliminary examination of modelled bloom timing shows the large interannual variability in the onset of the spring bloom, consistent with one-dimensional models of the region (Collins et al. (2009); Allen and Wolfe (2013)) and in the in-situ and satellite based observations (Suchy et al. (2019); Gower et al. (2013); Boldt et al. (2019)). The significant spatial variability seen here underscores the dynamical differences in environmental growing conditions in different regions of the Salish Sea, and provides an interesting direction for future research.

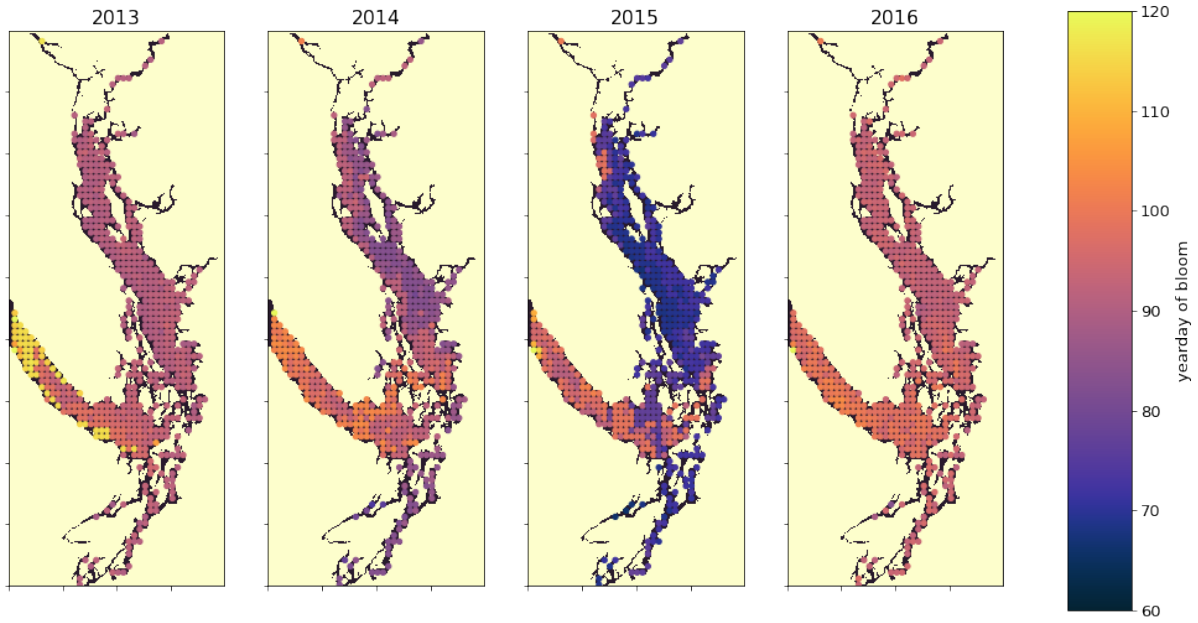

Figure 10. A spatial view of the onset of the spring bloom in the domain. Here the spring bloom is defined as the first peak in depth-integrated diatom biomass that is at least 30% of the maximum annual diatom biomass at that station. In all years, the spring bloom occurred earliest in the CSog and subsequently in the NSoG before reaching JdF with a variable delay.

4.4 Utility of Clustering Methods in the Context of High-Resolution Models

Our clustering approach identifies unambiguous regions of a complex coastal sea that exhibit distinct biological responses to disparate physical environments. These responses are not immediately obvious in time-averaged snapshots of the studied system. The simple machine learning technique used here enhances our way of looking at the problem - in this application, we are not using machine learning to predict unknown quantities, as is becoming common (e.g. Keppler et al. (2020)), but instead we are asking it to show us what is already there. Using this simple technique, we are able to draw objective boundaries between regions based on emergent structures in our data and significantly advance our intuition about the system. Cluster-based model evaluation may also be a very useful application of clustering techniques, as it has potential to diagnose how well a given model performs across different biophysical regimes.

The simplicity of the approach may have utility in numerous contexts. For instance, many characterizations of environmental regions rely on sparse data with large spatial biases. Objective clusters determined from regional models, with mechanistic under-pinnings, may be used to group sparse data. This approach allows clear characterization of complex systems. Furthermore, it may provide the necessary first step for machine learning studies that rely on well-organized training datasets to accurately predict target variables (e.g. Landschützer et al. (2013)). Resource and environmental management situations and optimal monitoring strategies may also benefit from a data-driven approach to regional definitions.

5 Conclusion

Our work applies a hierarchical clustering algorithm to four years of SalishSeaCast model output. We extract four factors relating to stratification and one relating to depth-integrated phytoplankton biomass, differentiated by functional group. We identify distinct regions of the model domain that exhibit contrasting wind and freshwater input dynamics, as well as regions of varying watercolumn-averaged vertical eddy diffusivity and halocline depth regimes. Similar spatial regionalizations in physical variables persist in all four analyzed years.

Similarly, we find distinct, interannually persisting, biological regions with phytoplankton biomass patterns that may be explained by patterns in the physical factors. In the NSoG, a deeper winter halocline and episodic summer mixing coincide with higher summer opportunist-type phytoplankton abundance, represented in the model by diatoms, and episodic fluctuations in phytoplankton biomass. In contrast, in the Fraser River stratified CSoG, shallower haloclines and stronger summer stratification coincide with more consistent biomass and high summer abundance of gleaner-type phytoplankton with slower growth rates, represented in the model as the flagellate functional group. While the biomass signals in the CSoG and NSoG suggest varying degrees of nutrient limitation, the JdF biomass signal suggests a light-limited physical regime. Furthermore, the cluster-based model evaluation suggests that JdF supports more biomass here than previously thought, due likely to a deeper growing layer. Our approach shows that stratification controls nutrient delivery and causes subtle structure in regional biological patterns, and demonstrates the utility of simple machine learning tools in extracting insight from large datasets in the context of oceanographic models.

Author contributions. TJ developed the cluster analysis with significant input from SEA and DI. EMO developed and tuned the biological model and developed the basis for the model evaluation. TJ wrote the manuscript with significant scientific input from EO, SEA, DI, and KS. TJ is supervised by SEA and DI.

Competing interests. The authors declare that no competing interests are present.

Code and data availability. Model namelists and postprocessing and analysis scripts are available from the associated GitHub repository (made public upon paper acceptance: https://github.com/tjarnikova/CLUSTER_OS). SalishSeaCast results are available from the SalishSeaCast ERDDAP server (https://salishsea.eos.ubc.ca/erddap/griddap/index.html). Observational data used in the model evaluation are available online from the Department of Fisheries and Oceans Canada (DFO): https://www.pac.dfo-mpo.gc.ca/science/oceans/data-donnees/index-eng.html. More information about SalishSeaCast can be found on the project web page (https://salishsea.eos.ubc.ca).

Acknowledgements. This work was funded by the Marine Environmental Observation, Prediction and Response (MEOPAR) Network of Canada (grant numbers 1.2, 7.2 and 65.9), as well as by the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant (grant number RGPIN-2016-03865) and the Government of Canada's Aquatic Climate Change Adaptation and Services Program (ACCASP). Computational resources were provided by Compute Canada for hindcast runs (grant numbers FT520, RRG2648 and RRG2969). The SalishSeaCast software environment was developed by Doug Latornell. TJ was partially funded by a personal Four-Year Fellowship through the University of British Columbia. TJ thanks Dr. Valentina Radic for the insightful data science course that inspired the concept of the study, as well as the Mesoscale Ocean and Atmospheric Dynamics group at the University of British Columbia for the collegial atmosphere and valuable feedback.

Appendix A: Model Evaluation

We evaluate the version of the SalishSeaCast biophysical model used in this clustering analysis regionally against available data from the Department of Fisheries and Oceans Canada (Ocean Sciences Division. Department of Fisheries and Oceans Canada (2020)), specifically; nitrate, dissolved silica, log-transformed chlorophyll, absolute salinity, and conservative temperature, along with the spread of locations and times of collection (Figures A1, A2; Tables A1, A2).

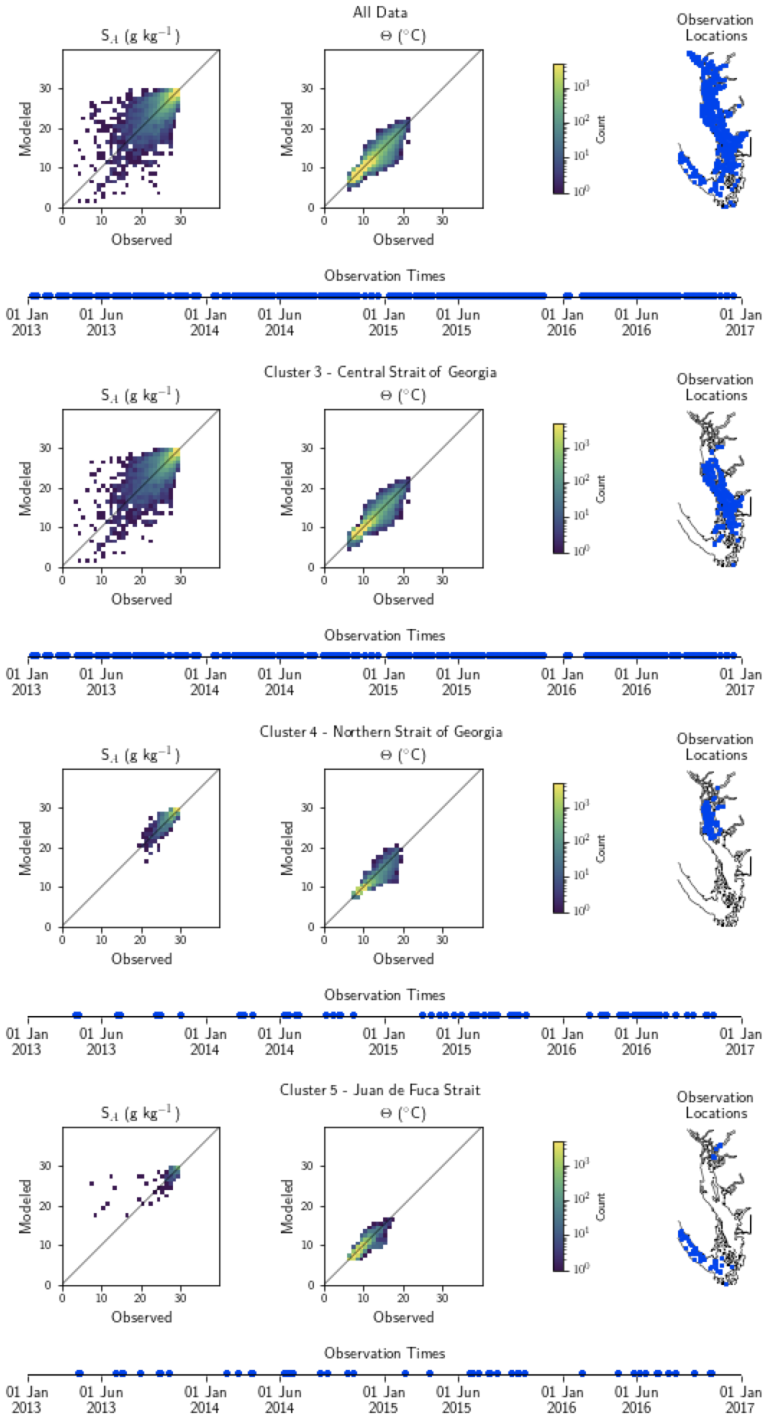

Figure A1. Model comparison with DFO CTD temperature and salinity data. The plots show modeled vs observed values for salinity and temperature for the entire model domain, as well as points matched only to the three major biological clusters - the Northern Strait of Georgia, the Central Strait of Georgia, and the Juan de Fuca Strait (cluster boundaries are specific to the year of observation). (Note that in some years, Bute Inlet clusters with the Juan de Fuca Strait). Because of the large amount of data available for comparison, a histogram view is presented. The timeline and rightmost panel display observation times and locations. Summary statistics corresponding to these plots are shown in Table A1.

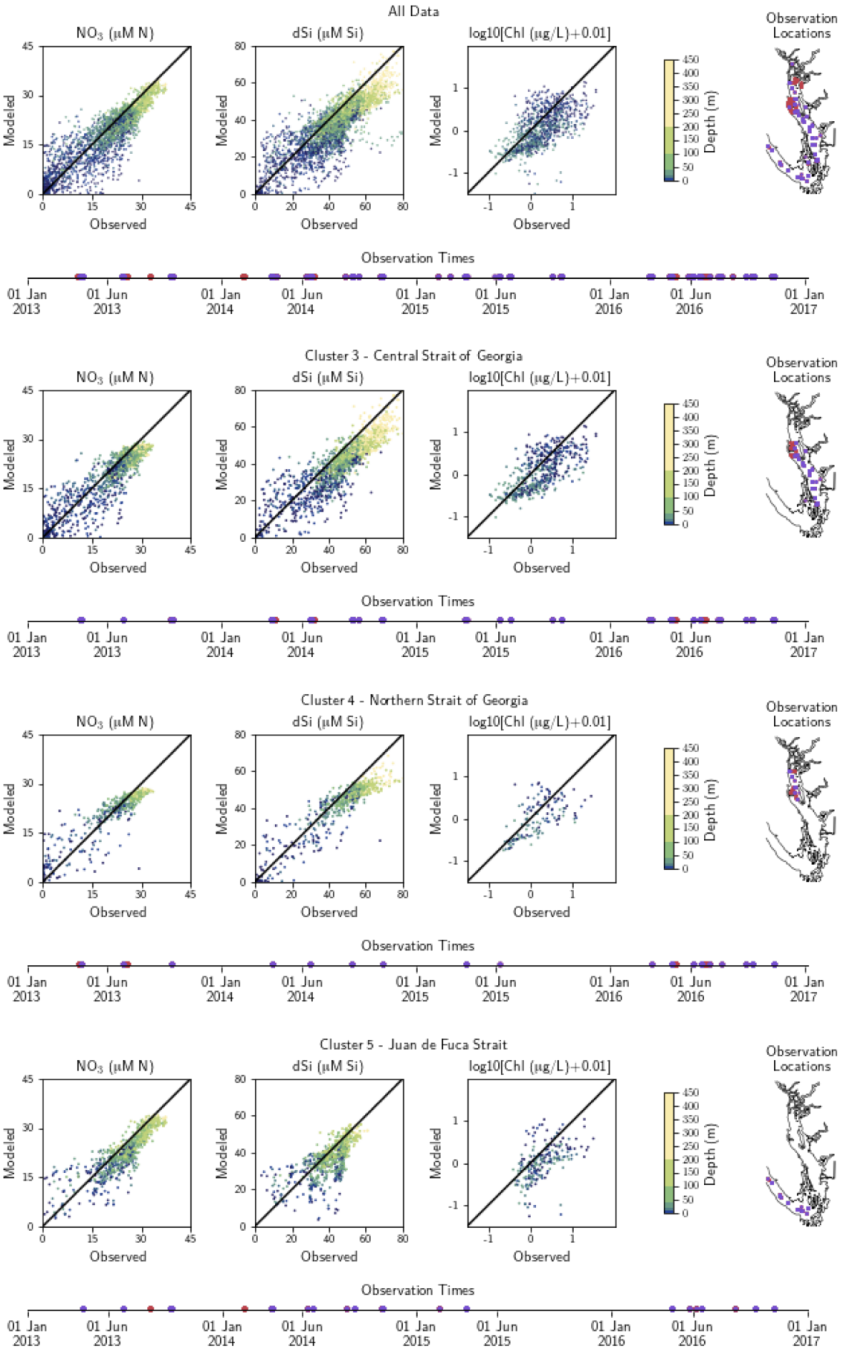

Figure A2. Model comparison with DFO nitrate, dissolved silica and log-transformed chlorophyll data. The plots show modeled vs observed values for nitrate, dissolved silica and log-transformed chlorophyll for the entire model domain, as well as points matched only to the three major biological clusters - the Northern Strait of Georgia, the Central Strait of Georgia, and the Juan de Fuca Strait (cluster boundaries are specific to the year of observation). The timeline and rightmost panel display observation times and locations. Stations with nutrients but no chlorophyll data are shown in red, while stations with observations of all three parameters are shown in purple. Summary statistics corresponding to these plots are shown in Table A2.

	metric	All data	Cluster 3 (CSoG)	Cluster 4 (NSoG)	Cluster 5 (JdF)
Temperature (°C)	N	502228	308314	56479	37858
	Model Mean	9.5	9.5	9.6	8.7
	Bias	0.01	0.044	-0.075	-0.068
	RMSE	0.47	0.44	0.45	0.51
	WSS	0.967	0.966	0.961	0.966
Salinity g/kg	N	502228	308314	56479	37858
	Model Mean	31	30	30	32
	Bias	0.046	0.067	0.15	-0.066
	RMSE	0.47	0.48	0.32	0.42
	WSS	0.967	0.960	0.970	0.971

Table A1. Summary statistics corresponding to the model-data comparison of temperature and salinity shown in Figure A1. Model bias is low compared to model means, and model bias and skill score do not vary significantly between biological clusters.

	metric	All data	Cluster 3 (CSoG)	Cluster 4 (NSoG)	Cluster 5 (JdF)
Nitrate	N	4732	2212	682	933
	Model Mean	21	22	22	23
	Bias	-2.0	-2.1	-0.94	-2.4
	RMSE	3.9	3.7	3.7	4.3
	WSS	0.94	0.97	0.95	0.90
Dissolved silica	N	4732	2212	682	933
	Model Mean	39	41	42	37
	Bias	-6.2	-7.0	-5.9	-4.2
	RMSE	9.7	9.7	9.1	8.57
	WSS	0.865	0.866	0.913	0.786
Chlorophyll (l10)	N	950	475	133	222
	Model Mean	-0.58	-0.69	-0.71	-0.55
	Bias	-0.23	-0.19	-0.17	-0.28
	RMSE	0.48	0.42	0.43	0.53
	WSS	0.712	0.786	0.757	0.599

Table A2. Summary statistics corresponding to the model-data comparison of biological variables shown in figure A2. Chlorophyll data are log-10 transformed. Model bias is low compared to model means and RMSE, and model bias and skill score do not vary significantly between biological clusters.

Appendix B: Changes to Biophysical Model Since First Publication

Several adjustments to the biological model have been made from the simulation described in Olson et al. (2020) to the present run. The most significant concerns the silicon cycle. The rate of silica dissolution was adjusted from $3.089 \times 10^{-6}\mathrm{s}^{-1}$ to 1.221×10^{-6} s^{-1}, and a bottom flux of silicon of 6.66×10^{-5} mmol $\mathrm{m}^{-2}\mathrm{s}^{-1}$ was added across the land-ocean interface below 250 m. The sinking rate of biogenic silicon was increased from 1.44×10^{-4} m s^{-1} to 3.108×10^{-4} m s^{-1}. The bottom reflection coefficient for biogenic silicon was increased from 0.8 to 0.92 and the reflection coefficient for diatoms was changed from 0.8 to 0. Additionally, the ratio of diatom silicon to nitrogen content was increased from 1.5 to 1.8 μmol Si:umol N.

Diatom growth parameters were adjusted slightly, with an increase of the optimal light level from 42 W m^{-2} to 45 W m^{-2}, an increase in dissolved silica half saturation constant from 1.2 to 2.2 μM Si, and a 1% decrease in maximum growth rate. The flagellate half saturation constant for ammonium increased from 0.1 to 0.2 μM N.

Several small adjustments were made to grazing rates, prey preferences, and predation threshold, primarily to decrease the minimum standing stock of phytoplankton and increase grazing by microzooplankton relative to mesozooplankton. Additionally, the seasonally varying mesozooplankton maximum grazing level was adjusted slightly, decreasing winter and mid-summer grazing rates and bringing the cycle forward by approximately 5 days. The western boundary and riverine nutrient concentrations have been updated. The namelists specifying these small adjustments are available from the paper's associated GitHub repository.

Appendix C: Supplementary figures

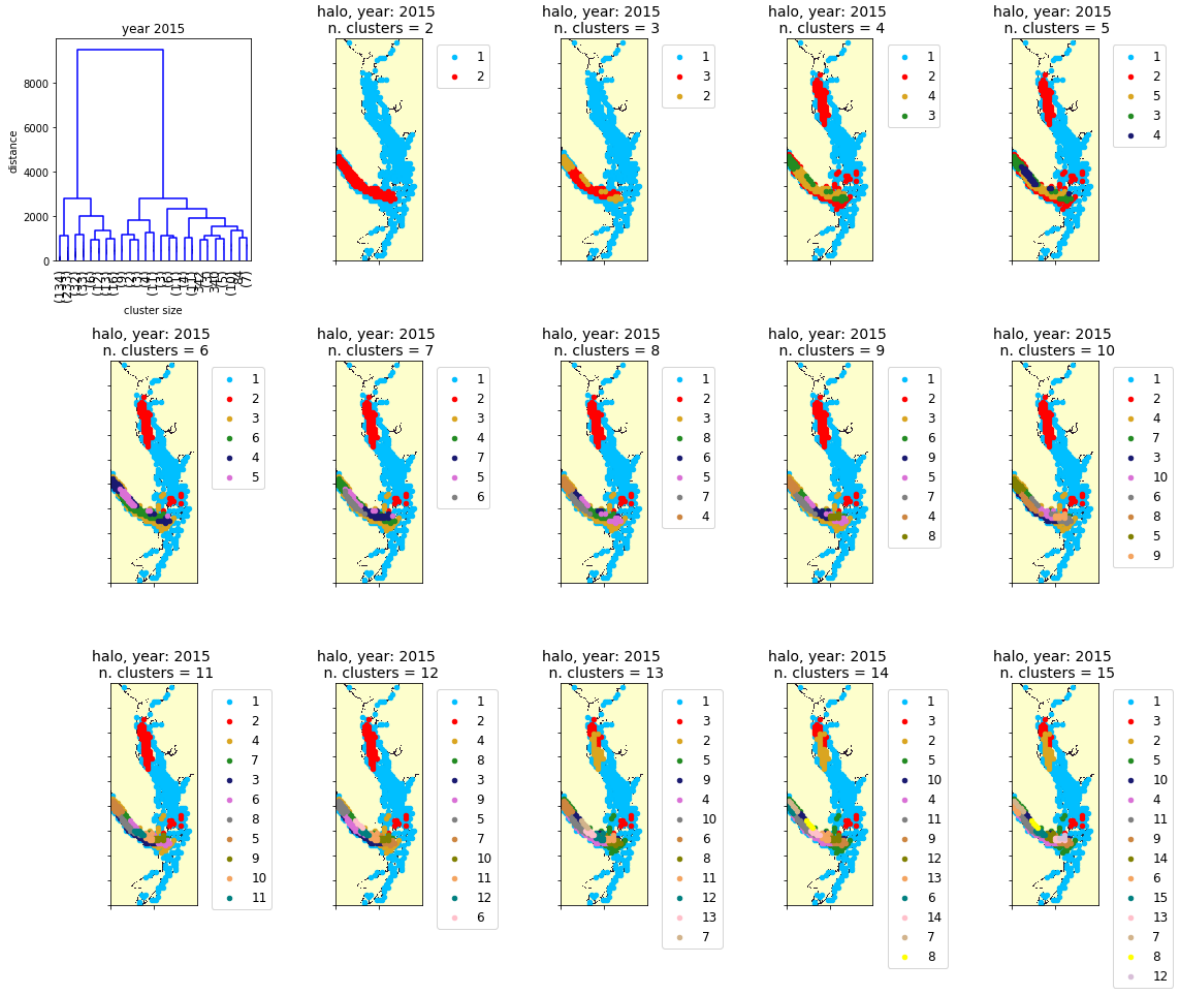

Figure C1. One example clustering output by Ward's method, for the annual halocline depth signal, year 2015 (see Sections 2.2-2.3).

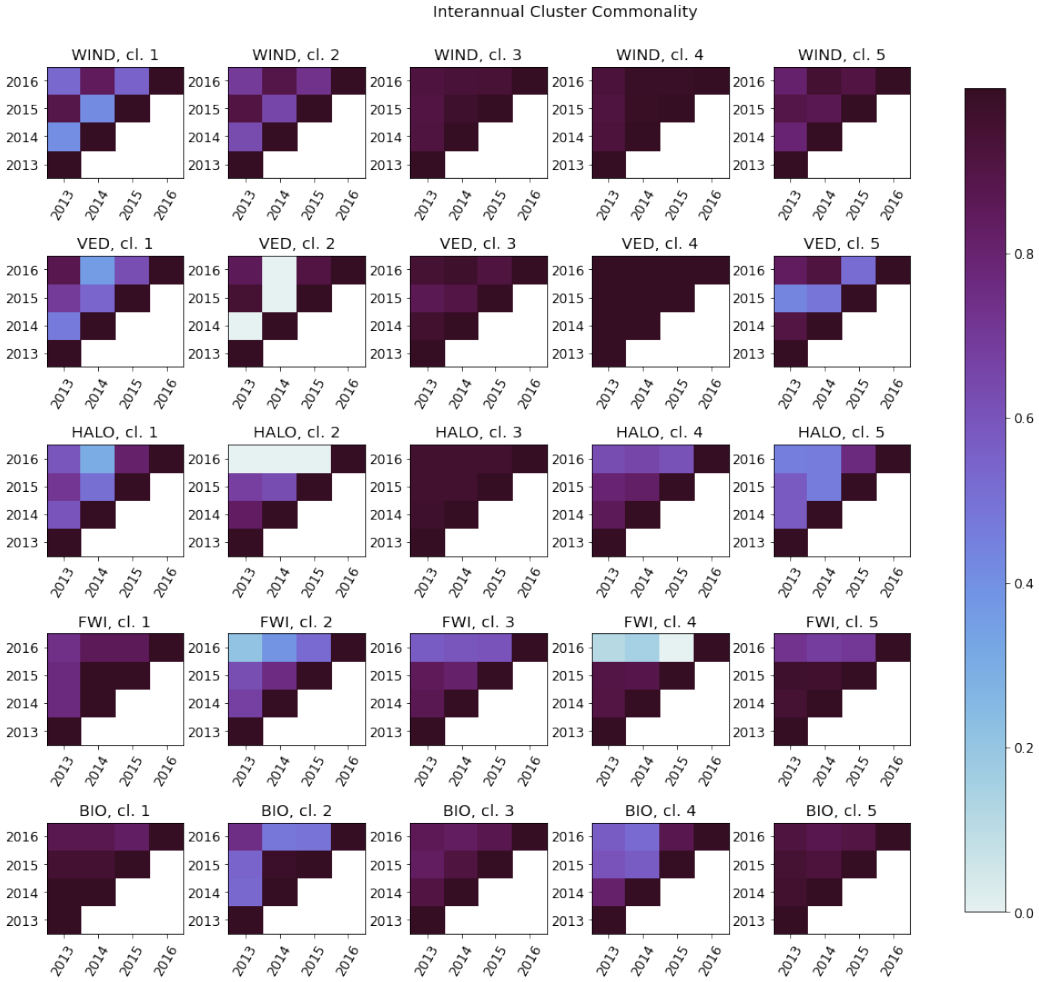

Figure C2. The interannual cluster commonality metric, measuring interannual cluster persistence for each factor. For any two clusters, cluster commonality varies from 0 (clusters of any size with no stations in common) to 1 (two clusters of equal size with all stations in common) and may be used to compare clusters of unequal sizes.

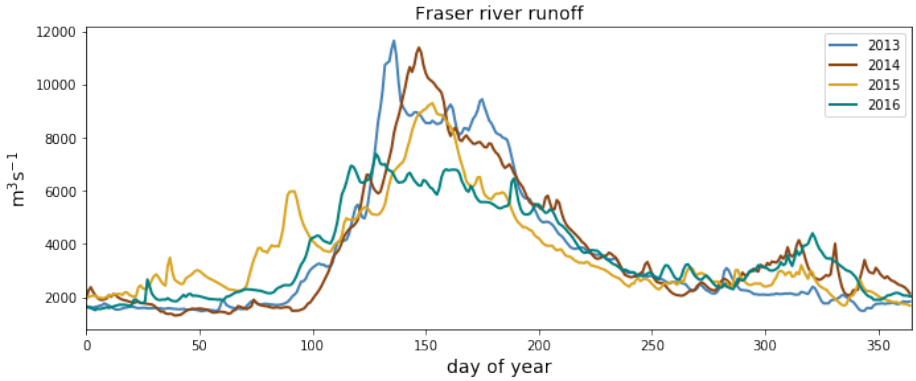

Figure C3. Fraser river flow at Hope, British Columbia for the four modelled years, as implemented in Soontiens and Allen (2017). Data from Environment and Climate Change Canada (https://wateroffice.ec.gc.ca/report/real_time_e.html?stn=08MF005, accessed June 2021).

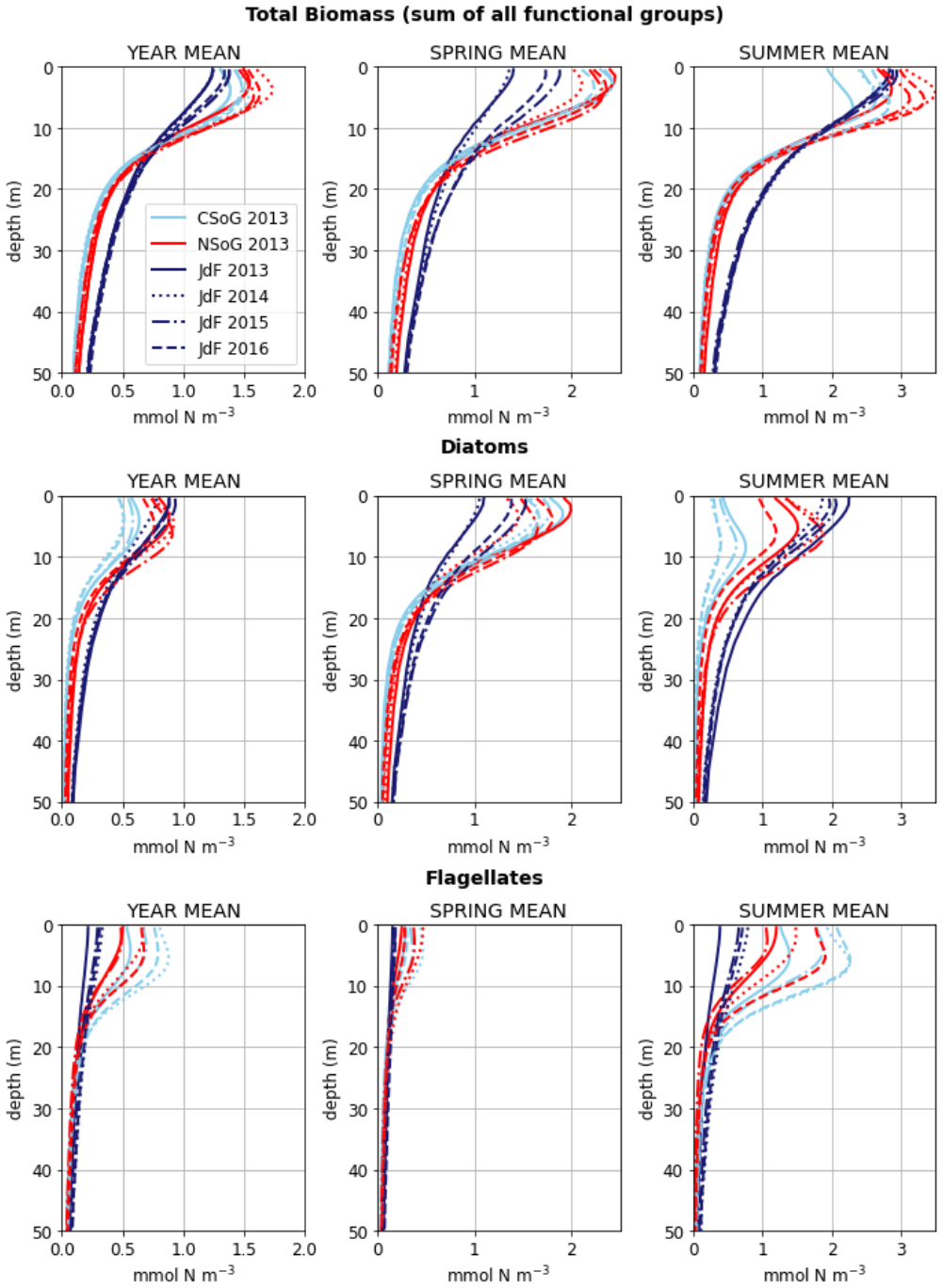

Figure C4. Mean depth profiles of phytoplankton biomass for the three main biological clusters (CSoG, NSoG, and JdF), for all four modeled years. Spring is defined as March-May, Summer is June-August, Autumn is September-November, and Winter is December-February.

References

Allen, S. and Wolfe, M.: Hindcast of the timing of the spring phytoplankton bloom in the Strait of Georgia, 1968–2010, Progress in Oceanography, 115, 6–13, 2013.

Anderies, J. M. and Beisner, B. E.: Fluctuating environments and phytoplankton community structure: a stochastic model, The American Naturalist, 155, 556–569, 2000.

Boldt, J. L., Thompson, M., Rooper, C. N., Hay, D. E., Schweigert, J. F., Quinn II, T. J., Cleary, J. S., and Neville, C. M.: Bottom-up and top-down control of small pelagic forage fish: factors affecting age-0 herring in the Strait of Georgia, British Columbia, Marine Ecology Progress Series, 617, 53–66, 2019.

Chandler, P. C., King, S. A., and Perry, R. I.: State of the physical, biological and selected fishery resources of Pacific Canadian marine ecosystems in 2016, Department of Fisheries and Oceans, 2016.

Cloern, J. E. and Dufford, R.: Phytoplankton community ecology: principles applied in San Francisco Bay, Marine Ecology Progress Series, 285, 11–28, 2005.

Collins, A. K., Allen, S. E., and Pawlowicz, R.: The role of wind in determining the timing of the spring bloom in the Strait of Georgia, Canadian Journal of Fisheries and Aquatic Sciences, 66, 1597–1616, 2009.

Crean, P.: A Numerical Model of Barotropic Mixed Tides Between Vancouver Island and the Mainland and its Relation to Studies of the Estuarine Circulation, Elsevier Oceanography Series, 23, 283–313, 1978.

Del Bel Belluz, J., Peña, M. A., Jackson, J. M., and Nemcek, N.: Phytoplankton Composition and Environmental Drivers in the Northern Strait of Georgia (Salish Sea), British Columbia, Canada, Estuaries and Coasts, pp. 1–21, 2021.

Deppe, R. W., Thomson, J., Polagye, B., and Krembs, C.: Predicting deep water intrusions to Puget Sound, WA (USA), and the seasonal modulation of dissolved oxygen, Estuaries and coasts, 41, 114–127, 2018.

Dutkiewicz, S., Follows, M. J., and Bragg, J. G.: Modeling the coupling of ocean ecology and biogeochemistry, Global Biogeochemical Cycles, 23, 2009.

Ebbesmeyer, C. C. and Barnes, C. A.: Control of a fjord basin's dynamics by tidal mixing in embracing sill zones, Estuarine and Coastal Marine Science, 11, 311–330, 1980.

Evans, W., Pocock, K., Hare, A., Weekes, C., Hales, B., Jackson, J., Gurney-Smith, H., Mathis, J. T., Alin, S. R., and Feely, R. A.: Marine CO_2 patterns in the northern salish sea, Frontiers in Marine Science, 5, 536, 2019.

Field, C. B., Behrenfeld, M. J., Randerson, J. T., and Falkowski, P.: Primary production of the biosphere: integrating terrestrial and oceanic components, science, 281, 237–240, 1998.

Fischer, H., List, E., Koh, R., Imberger, J., and Brooks, N.: Mixing in inland and coastal waters, Academic Press, 1979.

Follows, M. J., Dutkiewicz, S., Grant, S., and Chisholm, S. W.: Emergent biogeography of microbial communities in a model ocean, science, 315, 1843–1846, 2007.

Gentemann, C. L., Fewings, M. R., and García-Reyes, M.: Satellite sea surface temperatures along the West Coast of the United States during the 2014–2016 northeast Pacific marine heat wave, Geophysical Research Letters, 44, 312–319, 2017.

Giddings, S. N. and MacCready, P.: Reverse Estuarine Circulation Due to Local and Remote Wind Forcing, Enhanced by the Presence of Along-Coast Estuaries, Journal of Geophysical Research: Oceans, 122, 10 184–10 205, https://doi.org/https://doi.org/10.1002/2016JC012479, 2017.

Gower, J., King, S., Statham, S., Fox, R., and Young, E.: The Malaspina Dragon: a newly-discovered pattern of the early spring bloom in the Strait of Georgia, British Columbia, Canada, Progress in Oceanography, 115, 181–188, 2013.

Grover, J. P.: Resource competition in a variable environment: phytoplankton growing according to Monod's model, The American Naturalist, 136, 771–789, 1990.

Grover, J. P., HUDZIAK, J., and Grover, J. D.: Resource competition, vol. 19, Springer Science & Business Media, 1997.

Haigh, R. and Taylor, F.: Mosaicism of microplankton communities in the northern Strait of Georgia, British Columbia, Marine Biology, 110, 301–314, 1991.

Huisman, J., Sharples, J., Stroom, J. M., Visser, P. M., Kardinaal, W. E. A., Verspagen, J. M., and Sommeijer, B.: Changes in turbulent mixing shift competition for light between phytoplankton species, Ecology, 85, 2960–2970, 2004.

Keppler, L., Landschützer, P., Gruber, N., Lauvset, S., and Stemmler, I.: Seasonal Carbon Dynamics in the Near-Global Ocean, Global Biogeochemical Cycles, p. e2020GB006571, 2020.

Khangaonkar, T., Long, W., and Xu, W.: Assessment of circulation and inter-basin transport in the Salish Sea including Johnstone Strait and Discovery Islands pathways, Ocean Modelling, 109, 11–32, 2017.

Landschützer, P., Gruber, N., Bakker, D. C., Schuster, U., Nakaoka, S.-i., Payne, M. R., Sasse, T. P., and Zeng, J.: A neural network-based estimate of the seasonal to inter-annual variability of the Atlantic Ocean carbon sink, Biogeosciences, 10, 7793–7815, 2013.

LeBlond, P. H.: The Strait of Georgia: functional anatomy of a coastal sea, Canadian Journal of Fisheries and Aquatic Sciences, 40, 1033–1063, 1983.

Legendre, L.: Hydrodynamic control of marine phytoplankton production: the paradox of stability, in: Elsevier Oceanography Series, vol. 32, pp. 191–207, Elsevier, 1981.

Liu, J.: Evaluation of a NEMO model of the Strait of Georgia and insights into mixing and transport of the Fraser River plume, Master's thesis, University of British Columbia, Vancouver, British Columbia, https://open.library.ubc.ca/cIRcle/collections/ubctheses/24/items/1.0343600, 2014.

Longhurst, A., Sathyendranath, S., Platt, T., and Caverhill, C.: An estimate of global primary production in the ocean from satellite radiometer data, Journal of plankton Research, 17, 1245–1271, 1995.

MacCready, P., McCabe, R. M., Siedlecki, S. A., Lorenz, M., Giddings, S. N., Bos, J., Albertson, S., Banas, N., and Garnier, S.: Estuarine circulation, mixing, and residence times in the Salish Sea, Journal of Geophysical Research: Oceans, 126, e2020JC016 738, 2021.

Mackas, D. L. and Harrison, P. J.: Nitrogenous nutrient sources and sinks in the Juan de Fuca Strait/Strait of Georgia/Puget Sound estuarine system: assessing the potential for eutrophication, Estuarine, Coastal and Shelf Science, 44, 1–21, https://doi.org/https://doi.org/10.1006/ecss.1996.0110, 1997.

Madec, G., Bourdallé-Badie, R., Bouttier, P.-A., Bricaud, C., Bruciaferri, D., Calvert, D., Chanut, J., Clementi, E., Coward, A., Delrosso, D., et al.: NEMO ocean engine, 2017.

Mahara, N., Pakhomov, E., Dosser, H., and Hunt, B.: How zooplankton communities are shaped in a complex and dynamic coastal system with strong tidal influence, Estuarine, Coastal and Shelf Science, 249, 107 103, 2021.

Malick, M. J., Cox, S. P., Mueter, F. J., and Peterman, R. M.: Linking phytoplankton phenology to salmon productivity along a north–south gradient in the Northeast Pacific Ocean, Canadian Journal of Fisheries and Aquatic Sciences, 72, 697–708, 2015.

Mangiameli, P., Chen, S. K., and West, D.: A comparison of SOM neural network and hierarchical clustering methods, European Journal of Operational Research, 93, 402–417, 1996.

Masson, D.: Deep water renewal in the Strait of Georgia, Estuarine, Coastal and Shelf Science, 54, 115–126, 2002.

Masson, D. and Peña, A.: Chlorophyll distribution in a temperate estuary: The Strait of Georgia and Juan de Fuca Strait, Estuarine, Coastal and Shelf Science, 82, 19–28, 2009.

Maulik, U. and Bandyopadhyay, S.: Performance evaluation of some clustering algorithms and validity indices, IEEE Transactions on pattern analysis and machine intelligence, 24, 1650–1654, 2002.

Milbrandt, J. A., Bélair, S., Faucher, M., Vallée, M., Carrera, M. L., and Glazer, A.: The pan-Canadian high resolution (2.5 km) deterministic prediction system, Weather and Forecasting, 31, 1791–1816, https://doi.org/https://doi.org/10.1175/WAF-D-16-0035.1, 2016.

Moore, S. K., Mantua, N. J., Newton, J. A., Kawase, M., Warner, M. J., and Kellogg, J. P.: A descriptive analysis of temporal and spatial patterns of variability in Puget Sound oceanographic properties, Estuarine, Coastal and Shelf Science, 80, 545–554, 2008.

Moore-Maley, B. and Allen, S. E.: Wind-driven upwelling and surface nutrient delivery in a semi-enclosed coastal sea, Ocean Science, 18, 143–167, 2022.

Morrison, J., Foreman, M., and Masson, D.: A method for estimating monthly freshwater discharge affecting British Columbia coastal waters, Atmosphere-Ocean, 50, 1–8, 2012.

Nemcek, N., Hennekes, M., and Perry, I.: Seasonal dynamics of the phytoplankton community in the Salish Sea from HPLC, in: State of the Physical, Biological and Selected Fishery Resources of Pacific Canadian Marine Ecosystems in 2019, edited by Boldt, J. L., Javorski, A., and Chandler, P. C., chap. 39, pp. 169–173, Canadian Technical Report of Fisheries and Aquatic Sciences 3377, 2020.

Ocean Sciences Division. Department of Fisheries and Oceans Canada : Institute of Ocean Sciences Data Archive, http://www.pac.dfo-mpo.gc.ca/science/oceans/data-donnees/index-eng.html., 2020.

Olson, E. M., Allen, S. E., Do, V., Dunphy, M., and Ianson, D.: Assessment of Nutrient Supply by a Tidal Jet in the Northern Strait of Georgia Based on a Biogeochemical Model, Journal of Geophysical Research: Oceans, 125, e2019JC015 766, 2020.

Parsons, T., Stronach, J., Borstad, G., Louttit, G., and Perry, R.: Biological fronts in the Strait of Georgia, British Columbia, and their relation to recent measurements of primary productivity, Mar. Ecol. Prog. Ser, 6, 237–242, 1981.

Pawlowicz, R., Riche, O., and Halverson, M.: The circulation and residence time of the Strait of Georgia using a simple mixing-box approach, Atmosphere-Ocean, 45, 173–193, 2007.

Pawlowicz, R., Suzuki, T., Chappell, R., Ta, A., and Esenkulova, S.: Atlas of oceanographic conditions in the Strait of Georgia (2015-2019) based on the Pacific Salmon Foundation's Citizen Science Dataset, Canadian Technical Report of Fisheries and Aquatic Sciences, 3374, 2020.

Peña, M. A., Masson, D., and Callendar, W.: Annual plankton dynamics in a coupled physical–biological model of the Strait of Georgia, British Columbia, Progress In Oceanography, 146, 58–74, 2016.

Pike, R. G., Redding, T., Moore, R., Winkler, R., Bladon, K., et al.: Compendium of forest hydrology and geomorphology in British Columbia., Land Management Handbook-Ministry of Forests and Range, British Columbia, 2010.

Preikshot, D., Beamish, R. J., and Neville, C. M.: A dynamic model describing ecosystem-level changes in the Strait of Georgia from 1960 to 2010, Progress in Oceanography, 115, 28–40, 2013.

Richardson, A. J.: In hot water: zooplankton and climate change, ICES Journal of Marine Science, 65, 279–295, 2008.

Sonnewald, M., Dutkiewicz, S., Hill, C., and Forget, G.: Elucidating ecological complexity: Unsupervised learning determines global marine eco-provinces, Science Advances, 6, eaay4740, 2020.

Soontiens, N. and Allen, S. E.: Modelling sensitivities to mixing and advection in a sill-basin estuarine system, Ocean Modelling, 112, 17–32, 2017.

Soontiens, N., Allen, S. E., Latornell, D., Le Souëf, K., Machuca, I., Paquin, J.-P., Lu, Y., Thompson, K., and Korabel, V.: Storm surges in the Strait of Georgia simulated with a regional model, Atmosphere-Ocean, 54, 1–21, 2016.

St. John, M., Marinone, S., Stronach, J., Harrison, P., Fyfe, J., and Beamish, R.: A horizontally resolving physical–biological model of nitrate concentration and primary productivity in the Strait of Georgia, Canadian Journal of Fisheries and Aquatic Sciences, 50, 1456–1466, 1993.

Suchy, K. D., Le Baron, N., Hilborn, A., Perry, R. I., and Costa, M.: Influence of environmental drivers on spatio-temporal dynamics of satellite-derived chlorophyll a in the Strait of Georgia, Progress in Oceanography, 176, 102 134, 2019.

Sun, Q., Little, C. M., Barthel, A. M., and Padman, L.: A clustering-based approach to ocean model–data comparison around Antarctica, Ocean Science, 17, 131–145, 2021.

Sutton, J., Johannessen, S., and Macdonald, R.: A nitrogen budget for the Strait of Georgia, British Columbia, with emphasis on particulate nitrogen and dissolved inorganic nitrogen, Biogeosciences, 10, 7179–7194, https://doi.org/https://doi.org/10.5194/bg-10-7179-2013, 2013.

Sverdrup, H.: On conditions for the vernal blooming of phytoplankton, J. Cons. Int. Explor. Mer, 18, 287–295, 1953.

Thomson, R.: Oceanography of the British Columbia coast, Department of Fisheries and Oceans, Ottawa, 1981.

Thomson, R., Miha'ly, S., and Kulikov, E.: Estuarine versus transient flow regimes in Juan de Fuca Strait, Journal of Geophysical Research, 112, https://doi.org/10.1029/2006JC003925, 2007.

Umlauf, L. and Burchard, H.: A generic length-scale equation for geophysical turbulence models, Journal of Marine Research, 61, 235–265, 2003.

Ward Jr, J. H.: Hierarchical grouping to optimize an objective function, Journal of the American statistical association, 58, 236–244, 1963.

Wasser, S. K., Lundin, J. I., Ayres, K., Seely, E., Giles, D., Balcomb, K., Hempelmann, J., Parsons, K., and Booth, R.: Population growth is limited by nutritional impacts on pregnancy success in endangered Southern Resident killer whales (Orcinus orca), PLoS One, 12, e0179 824, 2017.

Wishart, D.: 256. Note: An algorithm for hierarchical classifications, Biometrics, pp. 165–170, 1969.

Yin, K., Goldblatt, R. H., Harrison, P. J., John, M. A. S., Clifford, P. J., and Beamish, R. J.: Importance of wind and river discharge in influencing nutrient dynamics and phytoplankton production in summer in the central Strait of Georgia, Marine Ecology Progress Series, 161, 173–183, 1997.