Report Card Analytics

NESP 3.2.5 WATER QUALITY METRIC

Author: Murray Logan

AIMS: Australia's tropical marine research agency

December 22, 2016

Australian Institute of Marine Science
PMB No 3 Townsville MC QLD 4810
PO Box 41775 Casuarina NT 0811
The UWA Oceans Institute (M096) Crawley WA 6009

This report should be cited as:

Enquires should be directed to:

Murray Logan
m.logan@aims.gov.au

DISCLAIMER

Revision History

Version		Title	Name	Date	Comments
1	Author	Dr	Murray Logan	December 22, 2016	
	Approved by				
2	Author				
	Approved by				
3	Author				
	Approved by				
4	Author				
	Approved by				

1 EXECUTIVE SUMMARY

- Guideline values are strictly intended to be applied to annually aggregated data. Similarly, where available, seasonal guideline values are intended to be applied to data aggregated within the wet and dry seasons.
- At the insistence of GBRMPA, these values must be applied to individual observations (rather than aggregations). For this reason, we have elected to refer to them as **thresholds** rather than guidelines.
- The spatial domain for these analyses is the entire GBR Marine Park along with Port regions extending from Cape York in the North to just North of Frazer Island in the South
- The GBR is further broken down into 24 **Zones** that represent each of six **Regions** (Cape York, Wet Tropics, Dry Tropics, Mackay Whitsunday, Fitzroy and Burnett Mary) and four **Water Bodies** (Enclosed Coastal, Open Coastal, Midshelf and Offshore).

Table 1: Water Quality Threshold values for each Measure in each Zone (Region/Water Body). Thresholds values are similar to annual Guideline values. Wet and Dry represent Wet and Dry season thresholds respectively. Direction of Failure indicates whether a values higher ('H') or lower ('L') than a Threshold would constitute an exceedence. Range From and Range To represent Thresholds for Measures that have a range of optimum values (such as dissolved oxygen or pH).

Measure	Units	Water Body	Region	Threshold			Direction of Faiure	Justification
				Annual	Dry	Wet		
chl	μgL^{-1}	Enclosed Coastal	Cape York	2.00	2.00	2.00	H	QLD WQ guidelines
chl	μgL^{-1}	Enclosed Coastal	Wet Tropics	2.00	2.00	2.00	H	There is no seasonal adjustment
chl	μgL^{-1}	Enclosed Coastal	Dry Tropics	2.00	2.00	2.00	H	
chl	μgL^{-1}	Enclosed Coastal	Mackay Whitsunday	2.00	2.00	2.00	H	
chl	μgL^{-1}	Enclosed Coastal	Fitzroy	2.00	2.00	2.00	H	
chl	μgL^{-1}	Enclosed Coastal	Burnett Mary	2.00	2.00	2.00	H	
chl	μgL^{-1}	Open Coastal	Cape York	0.45	0.32	0.63	H	GBRMPA WQ guidelines
chl	μgL^{-1}	Open Coastal	Wet Tropics	0.45	0.32	0.63	H	40% higher in summer, 30% lower in winter
chl	μgL^{-1}	Open Coastal	Dry Tropics	0.45	0.32	0.63	H	Here summer is taken as Wet Season
chl	μgL^{-1}	Open Coastal	Mackay Whitsunday	0.45	0.32	0.63	H	and winter is taken as Dry Season
chl	μgL^{-1}	Open Coastal	Fitzroy	0.45	0.32	0.63	H	
chl	μgL^{-1}	Open Coastal	Burnett Mary	0.45	0.32	0.63	H	
chl	μgL^{-1}	Midshelf	Cape York	0.45	0.32	0.63	H	GBRMPA WQ guidelines
chl	μgL^{-1}	Midshelf	Wet Tropics	0.45	0.32	0.63	H	40% higher in summer, 30% lower in winter
chl	μgL^{-1}	Midshelf	Dry Tropics	0.45	0.32	0.63	H	Here summer is taken as Wet Season
chl	μgL^{-1}	Midshelf	Mackay Whitsunday	0.45	0.32	0.63	H	and winter is taken as Dry Season
chl	μgL^{-1}	Midshelf	Fitzroy	0.45	0.32	0.63	H	
chl	μgL^{-1}	Midshelf	Burnett Mary	0.45	0.32	0.63	H	
chl	μgL^{-1}	Offshore	Cape York	0.40	0.28	0.56	H	GBRMPA WQ guidelines
chl	μgL^{-1}	Offshore	Wet Tropics	0.40	0.28	0.56	H	40% higher in summer, 30% lower in winter
chl	μgL^{-1}	Offshore	Dry Tropics	0.40	0.28	0.56	H	Here summer is taken as Wet Season
chl	μgL^{-1}	Offshore	Mackay Whitsunday	0.40	0.28	0.56	H	and winter is taken as Dry Season
chl	μgL^{-1}	Offshore	Fitzroy	0.40	0.28	0.56	H	
chl	μgL^{-1}	Offshore	Burnett Mary	0.40	0.28	0.56	H	
nap	mgL^{-1}	Enclosed Coastal	Cape York	25.00	25.00	25.00	H	QLD WQ guidelines
nap	mgL^{-1}	Enclosed Coastal	Wet Tropics	25.00	25.00	25.00	H	There is no seasonal adjustment and
nap	mgL^{-1}	Enclosed Coastal	Dry Tropics	15.00	15.00	15.00	H	values for CY and WT are not determined
nap	mgL^{-1}	Enclosed Coastal	Mackay Whitsunday	15.00	15.00	15.00	H	Suggest applying same ratio as for turbidity
nap	mgL^{-1}	Enclosed Coastal	Fitzroy	15.00	15.00	15.00	H	between CY/WT and others, i.e (15*10)/6=25
nap	mgL^{-1}	Enclosed Coastal	Burnett Mary	15.00	15.00	15.00	H	NAP is taken as = TSS in this context
nap	mgL^{-1}	Open Coastal	Cape York	2.00	1.60	2.40	H	GBRMPA WQ guidelines
nap	mgL^{-1}	Open Coastal	Wet Tropics	2.00	1.60	2.40	H	20% higher in summer, 20% lower in winter
nap	mgL^{-1}	Open Coastal	Dry Tropics	2.00	1.60	2.40	H	Here summer is taken as Wet Season
nap	mgL^{-1}	Open Coastal	Mackay Whitsunday	2.00	1.60	2.40	H	and winter is taken as Dry Season

...continued from previous page

Measure	Units	Water Body	Region	Threshold			Direction of Faiure	Justification
				Annual	Dry	Wet		
nap	mgL^{-1}	Open Coastal	Fitzroy	2.00	1.60	2.40	H	
nap	mgL^{-1}	Open Coastal	Burnett Mary	2.00	1.60	2.40	H	NAP is taken as = TSS in this context
nap	mgL^{-1}	Midshelf	Cape York	2.00	1.60	2.40	H	GBRMPA WQ guidelines
nap	mgL^{-1}	Midshelf	Wet Tropics	2.00	1.60	2.40	H	20% higher in summer, 20% lower in winter
nap	mgL^{-1}	Midshelf	Dry Tropics	2.00	1.60	2.40	H	Here summer is taken as Wet Season
nap	mgL^{-1}	Midshelf	Mackay Whitsunday	2.00	1.60	2.40	H	and winter is taken as Dry Season
nap	mgL^{-1}	Midshelf	Fitzroy	2.00	1.60	2.40	H	
nap	mgL^{-1}	Midshelf	Burnett Mary	2.00	1.60	2.40	H	NAP is taken as = TSS in this context
nap	mgL^{-1}	Offshore	Cape York	0.70	0.56	0.84	H	GBRMPA WQ guidelines
nap	mgL^{-1}	Offshore	Wet Tropics	0.70	0.56	0.84	H	20% higher in summer, 20% lower in winter
nap	mgL^{-1}	Offshore	Dry Tropics	0.70	0.56	0.84	H	Here summer is taken as Wet Season
nap	mgL^{-1}	Offshore	Mackay Whitsunday	0.70	0.56	0.84	H	and winter is taken as Dry Season
nap	mgL^{-1}	Offshore	Fitzroy	0.70	0.56	0.84	H	
nap	mgL^{-1}	Offshore	Burnett Mary	0.70	0.56	0.84	H	NAP is taken as = TSS in this context
ntu	NTU	Enclosed Coastal	Cape York	10.00	10.00	10.00	H	QLD WQ guidelines
ntu	NTU	Enclosed Coastal	Wet Tropics	10.00	10.00	10.00	H	There is no seasonal adjustment
ntu	NTU	Enclosed Coastal	Dry Tropics	6.00	6.00	6.00	H	
ntu	NTU	Enclosed Coastal	Mackay Whitsunday	6.00	6.00	6.00	H	Unclear as to why the CY/WT guidelines
ntu	NTU	Enclosed Coastal	Fitzroy	6.00	6.00	6.00	H	are higher than Southern regions
ntu	NTU	Enclosed Coastal	Burnett Mary	6.00	6.00	6.00	H	
ntu	NTU	Open Coastal	Cape York	1.50	1.20	1.80	H	No guideline available but turbidity needed if logger
ntu	NTU	Open Coastal	Wet Tropics	1.50	1.20	1.80	H	data is to be integrated.
ntu	NTU	Open Coastal	Dry Tropics	1.50	1.20	1.80	H	MMP used a correlation between TSS and NTU
ntu	NTU	Open Coastal	Mackay Whitsunday	1.50	1.20	1.80	H	(based on whole of GBR data) which is also used here
ntu	NTU	Open Coastal	Fitzroy	1.50	1.20	1.80	H	
ntu	NTU	Open Coastal	Burnett Mary	1.50	1.20	1.80	H	Applied 20% higher in summer, 20% lower in winter
ntu	NTU	Midshelf	Cape York	1.50	1.20	1.80	H	No guideline available but turbidity needed if logger
ntu	NTU	Midshelf	Wet Tropics	1.50	1.20	1.80	H	data is to be integrated.
ntu	NTU	Midshelf	Dry Tropics	1.50	1.20	1.80	H	MMP used a correlation between TSS and NTU
ntu	NTU	Midshelf	Mackay Whitsunday	1.50	1.20	1.80	H	(based on whole of GBR data) which is also used here
ntu	NTU	Midshelf	Fitzroy	1.50	1.20	1.80	H	
ntu	NTU	Midshelf	Burnett Mary	1.50	1.20	1.80	H	Applied 20% higher in summer, 20% lower in winter
ntu	NTU	Offshore	Cape York	1.00	0.80	1.20	H	No guideline available but turbidity needed if logger
ntu	NTU	Offshore	Wet Tropics	1.00	0.80	1.20	H	data is to be integrated.
ntu	NTU	Offshore	Dry Tropics	1.00	0.80	1.20	H	MMP used a correlation between TSS and NTU
ntu	NTU	Offshore	Mackay Whitsunday	1.00	0.80	1.20	H	(based on whole of GBR data) which is also used here

...continued from previous page

Measure	Units	Water Body	Region	Threshold			Direction of Faiure	Justification
				Annual	Dry	Wet		
ntu	NTU	Offshore	Fitzroy	1.00	0.80	1.20	H	
ntu	NTU	Offshore	Burnett Mary	1.00	0.80	1.20	H	Applied 20% higher in summer, 20% lower in winter
sd	m	Enclosed Coastal	Cape York	1.00	1.00	1.00	L	QLD WQ guidelines
sd	m	Enclosed Coastal	Wet Tropics	1.00	1.00	1.00	L	There is no seasonal adjustment
sd	m	Enclosed Coastal	Dry Tropics	1.50	1.50	1.50	L	
sd	m	Enclosed Coastal	Mackay Whitsunday	1.50	1.50	1.50	L	Unclear as to why the CY/WT guidelines
sd	m	Enclosed Coastal	Fitzroy	1.50	1.50	1.50	L	are lower than Southern regions
sd	m	Enclosed Coastal	Burnett Mary	1.50	1.50	1.50	L	
sd	m	Open Coastal	Cape York	10.00	10.00	10.00	L	GBRMPA WQ guidelines
sd	m	Open Coastal	Wet Tropics	10.00	10.00	10.00	L	There is no seasonal adjustment
sd	m	Open Coastal	Dry Tropics	10.00	10.00	10.00	L	
sd	m	Open Coastal	Mackay Whitsunday	10.00	10.00	10.00	L	
sd	m	Open Coastal	Fitzroy	10.00	10.00	10.00	L	
sd	m	Open Coastal	Burnett Mary	10.00	10.00	10.00	L	
sd	m	Midshelf	Cape York	10.00	10.00	10.00	L	GBRMPA WQ guidelines
sd	m	Midshelf	Wet Tropics	10.00	10.00	10.00	L	There is no seasonal adjustment
sd	m	Midshelf	Dry Tropics	10.00	10.00	10.00	L	
sd	m	Midshelf	Mackay Whitsunday	10.00	10.00	10.00	L	
sd	m	Midshelf	Fitzroy	10.00	10.00	10.00	L	
sd	m	Midshelf	Burnett Mary	10.00	10.00	10.00	L	
sd	m	Offshore	Cape York	17.00	17.00	17.00	L	GBRMPA WQ guidelines
sd	m	Offshore	Wet Tropics	17.00	17.00	17.00	L	There is no seasonal adjustment
sd	m	Offshore	Dry Tropics	17.00	17.00	17.00	L	
sd	m	Offshore	Mackay Whitsunday	17.00	17.00	17.00	L	
sd	m	Offshore	Fitzroy	17.00	17.00	17.00	L	
sd	m	Offshore	Burnett Mary	17.00	17.00	17.00	L	

2 AIMS IN SITU SAMPLES

The AIMS *in situ* data are processed in an identical manner to the data used in the MMP Water Quality report. Indeed, the data employed here, are the processed MMP data. The processing steps include:

- applying QAQC routines (such as limit of detection rules)
- expressing observations in reporting units
- averaging over duplicate samples
- coupling samples collected over consecutive days
- calculating depth weighted averages

Figures 1 and 2 respectively illustrate the spatial and temporal distribution of the AIMS *in situ* samples. Note, only some of the 24 Zones (Region/Water bodies) are represented in the AIMS *in situ* data. Figure 3 explores the spatio-temporal trends in the AIMS *in situ* data. Note data are grouped into MMP sub-regions purely for the purpose of evening up the number of reefs (Sites) plotted per row.

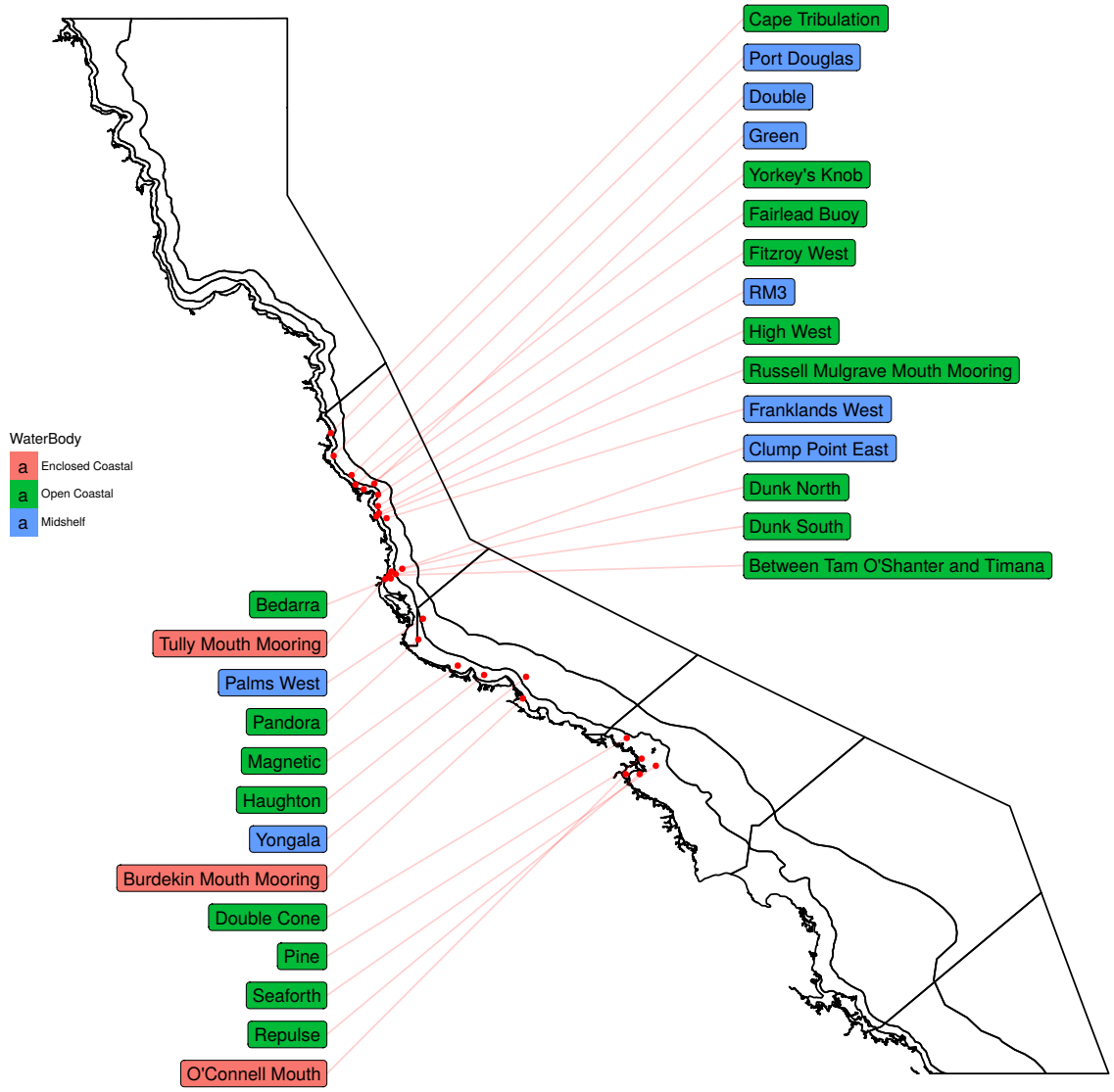


Figure 1: Map of AIMS in situ samples.

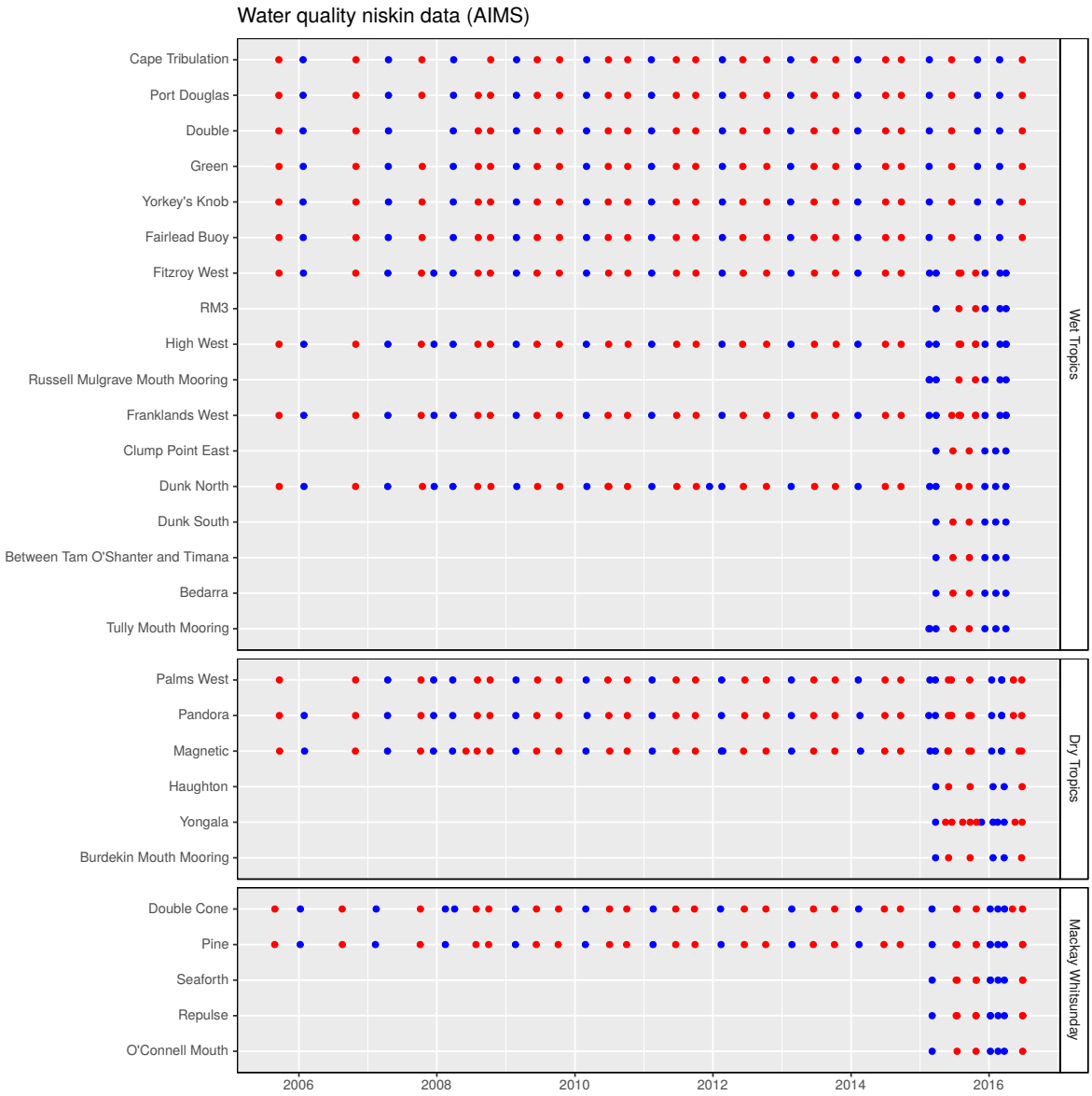


Figure 2: Temporal distribution of AIMS in situ samples. Red and Blue symbols represent samples collected in Dry and Wet seasons respectively.

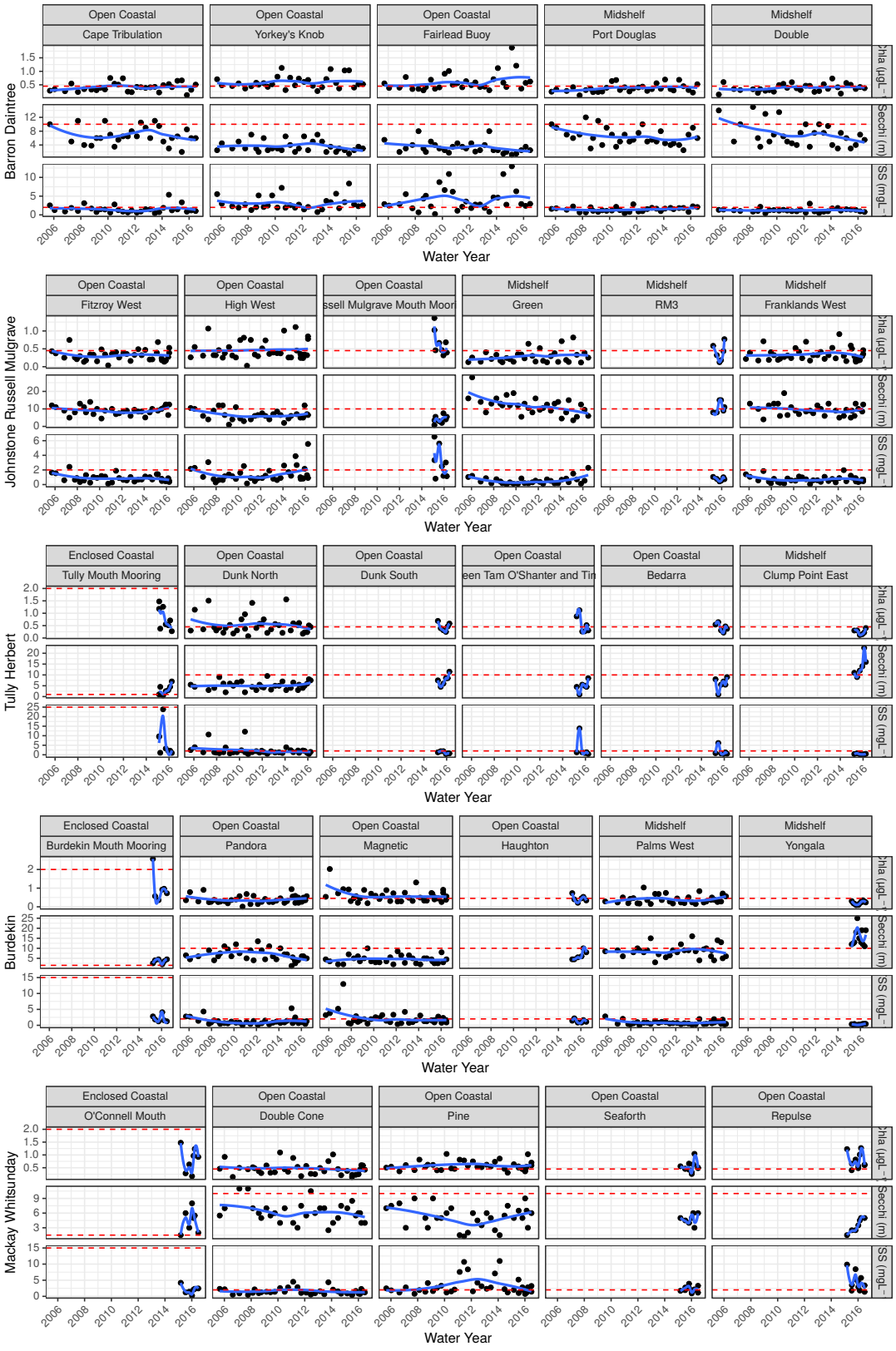


Figure 3: Time series of AIMS in situ samples. GAM smoothers applied and red dashed line represents threshold value. Note, Total Suspended Solids and Secchi Depth labelled as nap and sd purely for convenience in applying thresholds.

2.1 Explore simple aggregations

In order to aid and guide the interpretive transition of data through various indexation and aggregation stages, I will generate various summary figures.

- Simple annual means (aggregating to the level of water year) (Fig. 4 on page 12)
- Simple Zone/annual means (aggregating to the level of water year for each Zone) (Fig. 5 on page 13)

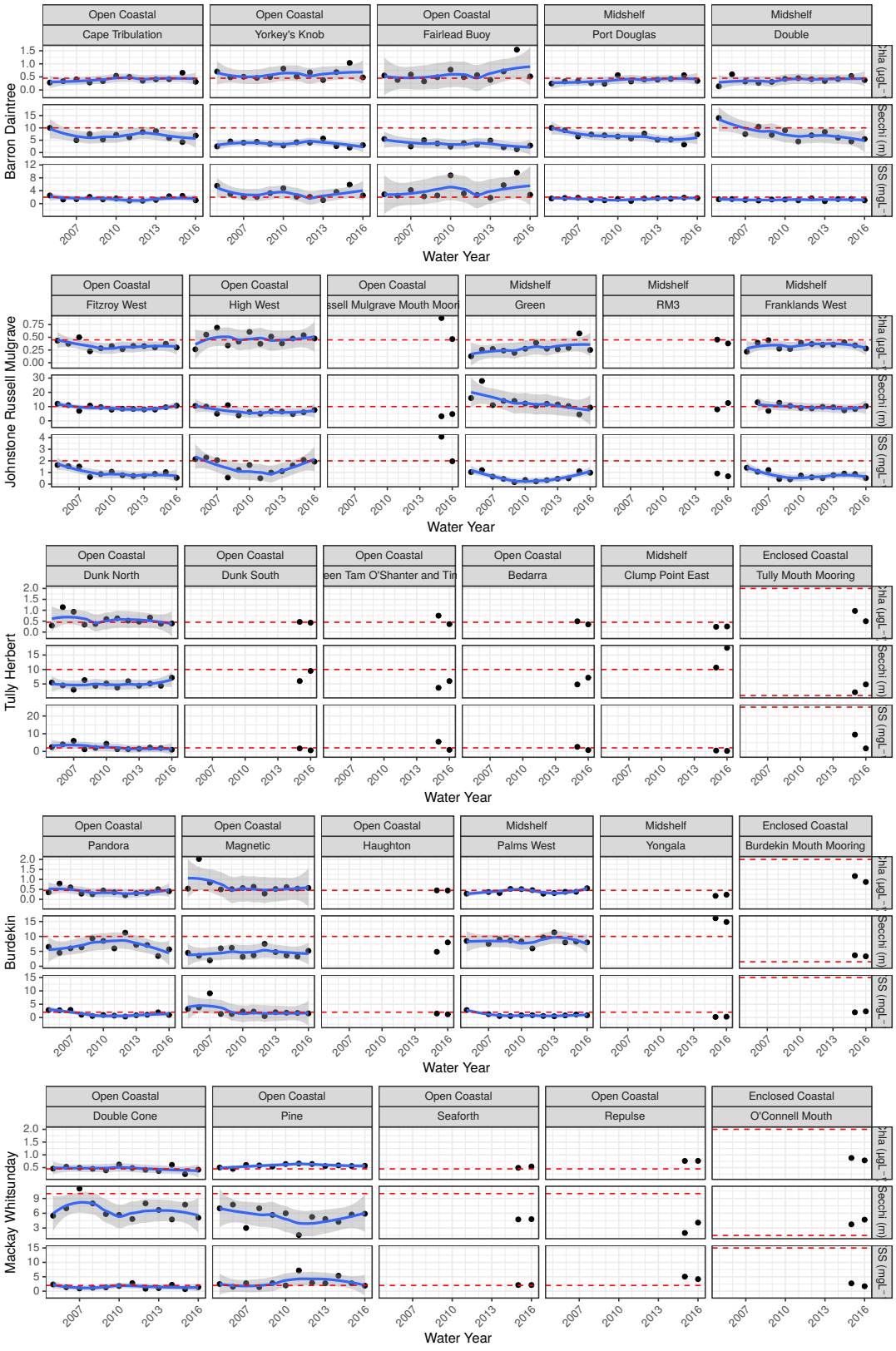


Figure 4: Time series (annual averages) of AIMS in situ samples. GAM smoothers applied and red dashed line represents threshold value. Note, Total Suspended Solids and Secchi Depth labelled as nap and sd purely for convenience in applying thresholds.

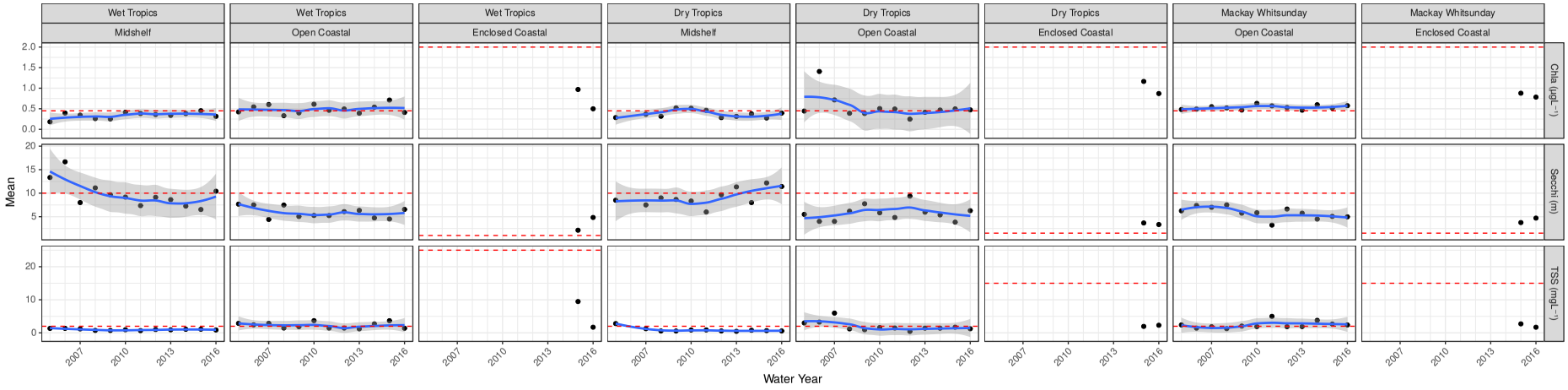


Figure 5: Time series (annual/region averages) of AIMS in situ samples. GAM smoothers applied.

2.2 Indices

Water Quality indices (which are standardized measures of condition) are typically expressed relative to a guideline, threshold (see Table 1 on page 3) or benchmark. Of the numerous calculation methods available, those that take into account the distance from the guideline (i.e. incorporate *difference-to-reference*) rather than simply an indication of whether or not a guideline value has been exceeded are likely to retain more information as well as being less sensitive to small changes in condition close to the guidelines.

The challenging aspect of distance (or amplitude) based index methodologies is that what constitutes a large deviation from a benchmark depends on the scale of the measure. For example, a deviation of 10 units might be considered relatively large of turbidity (NTU) or salinity (ppt), yet might be considered only minor for the Chlorophyll-a ($\mu g/L$). In order to combine a range of such metrics together into a meaningful index, the individual scores must be expressed on a common scale. Whilst this is automatically the case for Binary compliance, it is not necessarily the case for distance based indices.

Table 2 describes and compares the formulations and response curves of the Binary compliance method as well as a number of amplitude (distance based) indexing methods.

The Modified Amplitude and Logistic Modified Amplitude are both based on a base 2 logarithm of the ratio of observed values to the associated be benchmark (see Table 2). This scale ensures that distances to the benchmark are symmetric (in that a doubling and halving equate to the same magnitude - yet apposing sign). Furthermore, the logarithmic transformation does provide some inbuilt capacity to accommodate log-normality (a common property of measured values).

By altering the sign of the exponent, the Modified Amplitude methods can facilitate stressors and responses for which a failure to comply with a benchmark would be either above or below the benchmark (e.g. NTU vs Secchi depth). Further modifications can be applied to accommodate measures in which the benchmark represents the ideal and deviations either above or below represent increasingly poorer conditions (e.g. pH and dissolved oxygen).

The raw Modified Amplitude scores are relatively insensitive to small fluctuations around a benchmarks and sensitivity increases exponentially with increasing distance to the benchmark. The resulting scores can take any value in the real line $[-\infty, \infty]$ and hence are not bounded[1] There are two broad approaches to scaling (see Table 2):

a. Capping and scaling: The log_2 scale can be capped to a range representing either a constant extent of change (e.g. twice and half the benchmark - a cap factor of 2) or else use historical quantiles (10th and 90th percentiles) to define the upper and lower bounds to which to cap the scale. Note historical quantiles are unavailable for the current application. Thereafter, either can be scaled to the range [0,1] via a simple formula (see Table 2 III.Scaled).

b. Logistic Modified Amplitude: By expressing the scores on a logistic scale, the range of scores can be automatically scaled to range [0,1]. Moreover, this method allows the shape of the response curve to be customized for purpose. For example, the relative sensitivity to changes close or far from the benchmarks can be altered by a tuning parameter.

Rather than aggregating across sites before calculating indices, we would suggest that indices should be calculated at the site level. This is particularly important when different measures are measured at different sites. Spatial variability can be addressed via the use of a bootstrapping routine (see below). We

[1] Unbounded indices are difficult to aggregate, since items that have very large magnitude scores will have more influence on the aggregation than those items with scores of smaller magnitude. Furthermore, unbounded scores are difficult to convert into alphanumeric Grades. Consequently, the Scores need to be scaled before they can be converted to alphabetical grading scale.

would recommend that measurements collected throughout the reporting year be aggregated together into a single annual value. This is primarily because most water quality guidelines pertain specifically to annual averages rather than single time samples. Although it is possible to incorporate uncertainty due to temporal variability, the low sparse temporal frequency of sample collection is likely to yield uncertainty characteristics that will swamp the more interesting spatial sources of uncertainty.

A useful metric for comparing the sensitivity of one indexing method over another is to take some representative longitudinal data and calculate indices based on the actual data as well as data that introduces progressively more noise.

Table 2: Formulations and example response curves for a variety of indicator scoring methods that compare observed values (x_i) to associated benchmark, guidelines or references values ($benchmark_i$ and dashed line). The Scaled Modified Amplitude Method can be viewed as three Steps: I. Initial Score generation, II. Score capping (two alternatives are provided) and III. Scaling to the range [0,1]. The first of the alternative capping formulations simply caps the Scores to set values (on a log_2 scale), whereas the second formulation (Quantile based, where $Q1$ and $Q2$ are quantiles) allows guideline quantiles to be used for capping purposes. Dotted lines represent capping boundaries. In the Logistic Scaled Amplitude method, T is a tuning parameter that controls the logistic rate (steepness at the inflection point). For the purpose of example, the benchmark was set to 50.

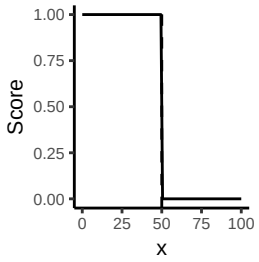
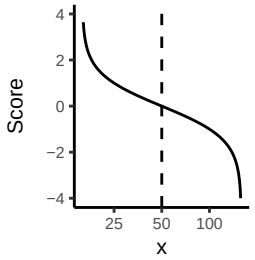
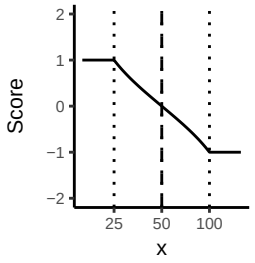
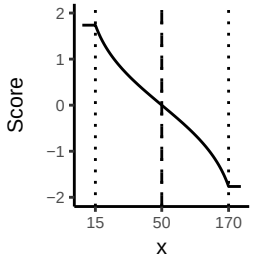
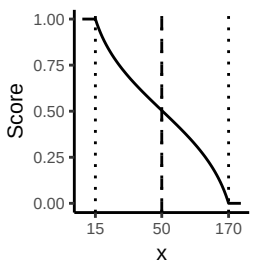
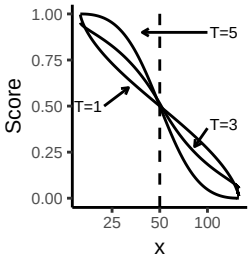
Method	**Formulation**	**Response curve**
Binary compliance	$Score_i = \begin{cases} 1 & \text{if } x_i \leq benchmark_i \\ 0 & \text{if } x_i \text{ else} \end{cases}$	
Modified Amplitude I. Raw (MAMP)	$Score_i = \begin{cases} log_2(\frac{x_i}{benchmark_i})^{-1} & \text{if} > benchmark_i = \text{fail} \\ log_2(\frac{x_i}{benchmark_i})^{1} & \text{if} < benchmark_i = \text{fail} \end{cases}$	
II. Fixed caps (-1,1)	$Score_i = \begin{cases} log_2(-1) & \text{if } Score_i < -1 \\ log_2(1) & \text{if } Score_i > 1 \\ Score_i & otherwise \end{cases}$	
II. Quantile based caps	$Score_i = \begin{cases} log_2(\frac{Q1}{benchmark_i})^{-1} & \text{if } x_i < Q1 \\ log_2(\frac{Q2}{benchmark_i})^{1} & \text{if } x_i > Q2 \\ Score_i & otherwise \end{cases}$	
III. Scaled	$Score_i = \frac{Score_i - min(Score_i)}{max(Score_i) - min(Score_i)}$	

Table 2: Report Card indexing methods, continued

Method	Formulation	Response curve
Logistic Scaled Modified Amplitude	Raw $Score_i = \begin{cases} log_2(\frac{x_i}{benchmark_i})^{-1} & \text{if} > benchmark_i = \text{fail} \\ log_2(\frac{x_i}{benchmark_i})^{1} & \text{if} < benchmark_i = \text{fail} \end{cases}$ $Score_i = \frac{1}{1+e^{Score_i \cdot -T}}$	

2.3 Hierarchical aggregation

To facilitate the integration of additional input Measures into the report card scores (such as additional Physical or Chemical), or even additional Sub-indicators (such as sediment metals, aquaculture yields etc), we can defined a hierarchical structure in which Measures (such as Chlorpohyll-a, NOx, sediment aluminum and yield etc) are nested within appropriate Sub-indicators. In turn, these Sub-indicators are nested within Indicators.

By progressively abstracting away the details of the Measures and Sub-indicators, a more focused narrative can be formulated around each level of the hierarchy. For example, when discussing the current state (and trend in state) of the Water Quality Indicator, rather than needing to discuss each individual constituent of Water Quality, high-level Grades are available on which to base high-level interpretations. More detailed explorations are thence revealed as required by exploring the Grades at progressively finer scales of the hierarchy. Moreover, the hierarchical structure offers great redundancy and thus flexibility to add, remove and exchange individual measures.

Similar arguments can be made for a spatial hierarchy in which Sites are nested within Zones which in turn are nested within the Whole Bay.

The purpose of aggregation is to combine together multiple items of data. For Nesp 3.2.5, the report card is informed by a triple hierarchical data structure in which Daily observations are nested within Seasonal and Annual aggregates, Measures are nested within Sub-indicators which are nested in Indicators and Sites are nested within Zones (see Figure 6).

Although the triple hierarchy (temporal, Spatial and Measurement), does offer substantial redundancy and power advantages, it also introduce the complexity of how to combine the hierarchies into a single hierarchical aggregation schedule. Table 3 (a fabricated example), illustrates this complexity for aggregating across Spatial and Measure scales when data availability differs. This simple example demonstrates how different aggregation schedules can result in different Zone Indicator scores:

- calculating Zone 1 Indicator Score as the average of the Site level Water Quality Scores prioritizes that the Zone 1 Indicator Score should reflect the average of the Water Quality Indicator Scores for the Site. This routine will bias the resulting Zone 1 Water Quality Indicator Score towards Sub-indicators represented in more Sites. The current MMP sampling design is unbalanced (some Zones have more Sites than others and not all Measures are observed in all Sites), and there is no guarantee that the design will be maintained over time. If for example, Chemical Measures were not available for certain Zones, then the Whole Bay Water Quality Indicator Score will be biased towards Water Clarity Sub-indicators.
- calculating Zone 1 Water Quality Indicator Score as the average of the Zone 1 level Sub-indicator Scores prioritizes equal contributions of Sub-indicators to the Indicator Score at the expense of being able to relate Zone 1 Scores to the corresponding Site Scores.

The above becomes even more complex when the temporal dimension is include..

An additional complication is how the different hierarchies integrate together. Specifically, what level of data should be aggregated first and at what point do the aggregations of one hierarchy feed into other hierarchies. For example, should observations first be aggregated from Daily to Seasonal or Annual, then aggregated from Site level to Zone level and then finally aggregated from Measure to Indicator? Some possible configurations are presented in Figure 7.

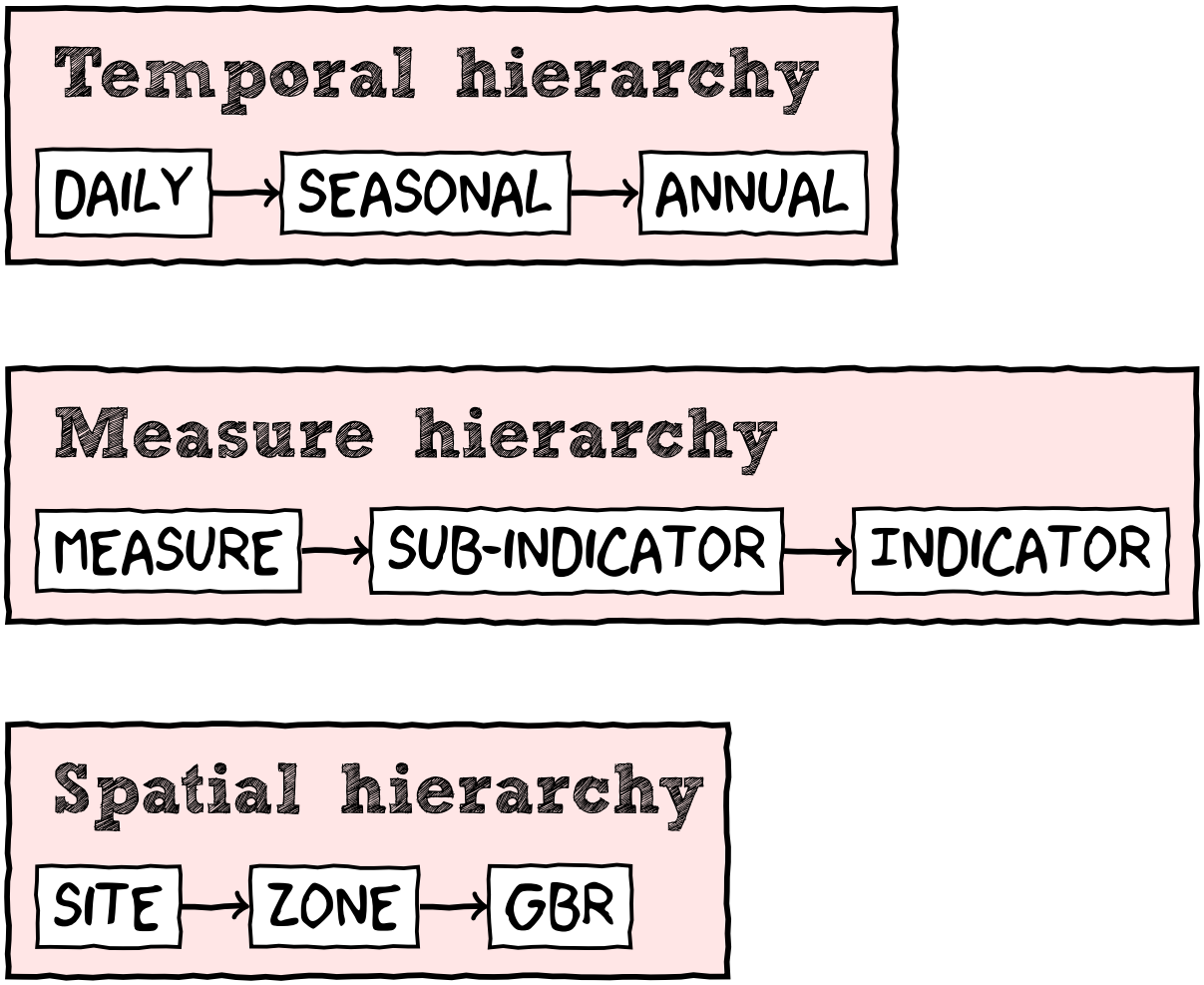


Figure 6: Temporal, measure and spatial aggregation hierarchy

Table 3: Fabricated illustration of the discrepancies between total means (i.e. Zone 1 Indicator Score) generated from row means (Site Sub-indicator Scores) and column means (Zone 1 Sub-indicator Scores).

	Sub-indicators		
Site	**Water Clarity**	**Nutrients**	**Indicator**
1	5	2	3.50
2	6		6.00
3	6	4	5.00
Zone 1	5.67	3.00	X

If X (mean) is calculated from the three row means = 4.83
If X (mean) is calculated from the two column means = 4.33

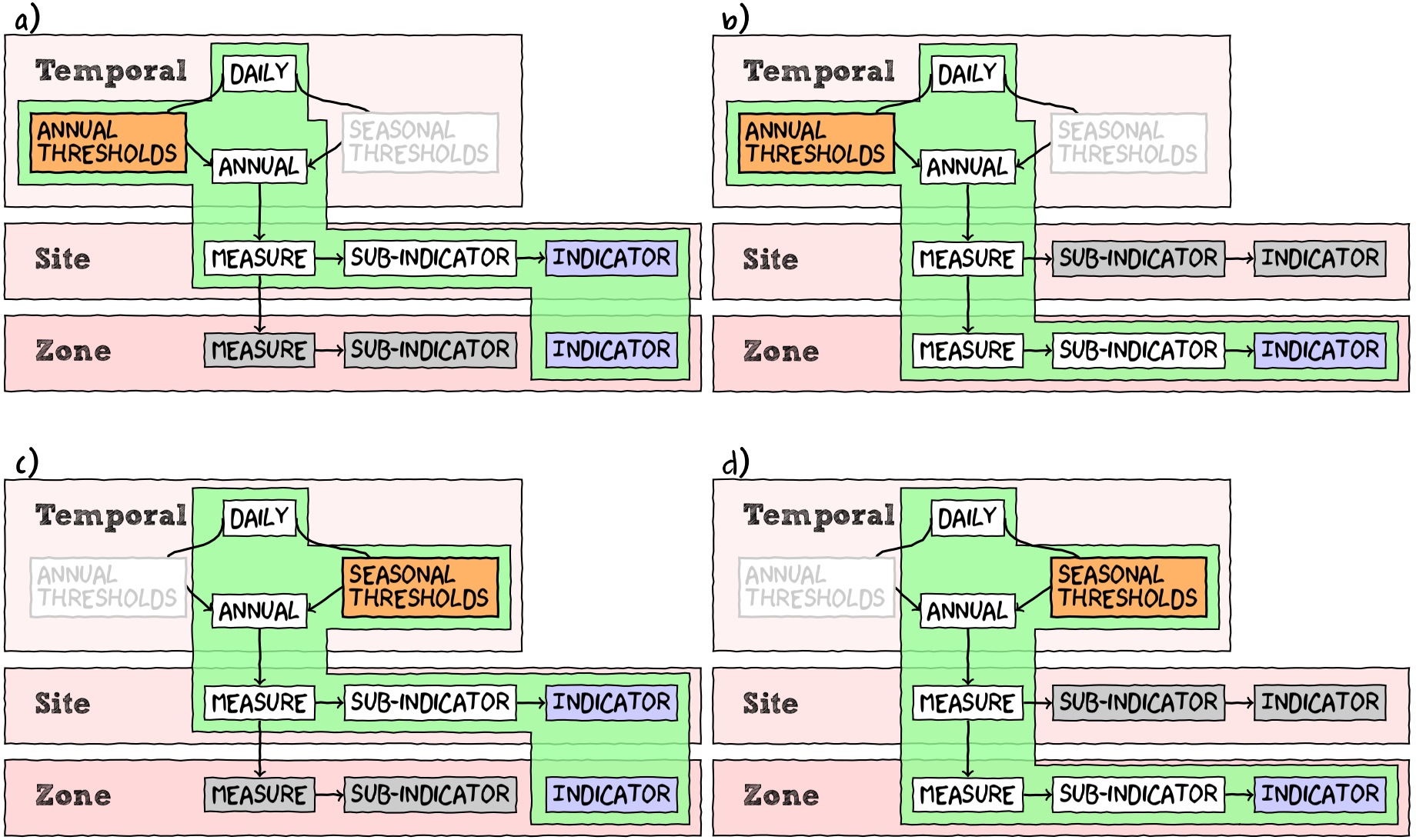


Figure 7: Schematic illustrating four possible aggregation routines through the combination of Temporal (Daily, Seasonal and Annual), Spatial (Site, Zone) and Measure (Measure, Sub-indicator, Indicator) nodes of the triple hierarchical aggregation routine associated with the GBR Report Card. Aggregation directions between nodes are signified by arrows and the main aggregation pathway through the routines is illustrated by the green polygon.

To maximize information retention throughout a series of aggregations, it is preferable to aggregate distributions rather than single properties of those distributions (such as means). The simplest way to perform a hierarchy of aggregations is to interactively calculate the means (or median) of items (means of means etc). At each successive aggregation level only very basic distributional summaries (such as the mean and perhaps standard deviation) are retained, the bulk of upstream information is lost. Alternatively, more complex methods that involve combining data or probability distributions can be effective at aggregating data in a way that propagates rich distributional properties throughout a series of aggregations.

Importantly, if the purpose of aggregation is purely to establish a new point estimate of the combined items, a large variety of methods essentially yield the same outcomes. On the other hand, if the purpose of aggregation is also to propagate a measure of uncertainty or confidence in the point estimate through multiple hierarchical levels of aggregation (as is the case here), then the different methodologies offer differing degrees of flexibility and suitability.

Hierarchical aggregations are essentially a series of steps that sequentially combine distributions (which progressively become more data rich). The resulting distribution formed at each step should thereby reflect the general conditions typified by its parent distributions and by extension, each of the distributions higher up the hierarchy.

Numerous characteristics can be estimated from a distribution including the location (such as mean and median) and scale (such as variance and range). For the current project, the mean and variance were considered the most appropriate[2] distributional descriptions and from these estimates Grades and measures of confidence can be respectively derived. Hence the numerical summaries (mean and variance) at any stage of the hierarchical aggregation are a byproduct rather than the sole property of propagation.

2.3.1 Bootstrap aggregation

Although some of the items to be aggregated together might initially comprise only a few values (or even a single value), it is useful to conceptualize them as continuous distributions. For example, when aggregating multiple *Measures* (such as all Water Quality Chemicals) together to generate a (*Site* level) *Sub-indicator* average, each *Measure* in each *Site* can be considered a distribution comprising the single Score for that *Measure*. Aggregation then involves combining together the multiple distributions into a single amalgam (by adding the distributions together, see Figure 8). Similarly, when aggregating at the *Indicator* level across *Site* to generate *Zone* summaries for each *Indicator*, *Site* distributions are respectively added together to yield a single distribution per *Zone*.

If the distributions being aggregated are all proportional distributions (e.g.~density distributions), adding them altogether is trivially simple. However, if, rather than actual distributions, the items to be aggregated are actually just small collections of values (as is the case for many of the discrete Measures here) or even large, yet unequally populous collections of values (as could be the case for Continuous Flow Monitoring with missing or suspect observations), then simply aggregating the distributions together will result in amalgams that are weighted according to the size of the collections (larger collections will have more influence). For example, if we were aggregating together three *Zones* (to yield Whole Bay estimates), one of which comprised twice as many Sites, simple aggregation of distributions would result in a distribution that was more highly influenced by the Zone with the more *Sites*. Similarly, when aggregating from the level of Sub-indicator to the level of Indicator, the resulting Indicator would be biased towards the Sub-indicator with the most Measures. Whilst this may well be a useful property (e.g.~stratified aggregation), it may also be undesirable.

[2]The aggregations typically involve some Measures with a small number of unique observations (and thus indices) and thus means and variances provide greater sensitivity than medians and ranges. Moreover, the indexing stage effectively removes outliers and standardizes the scale range thereby reducing the need for robust estimators.

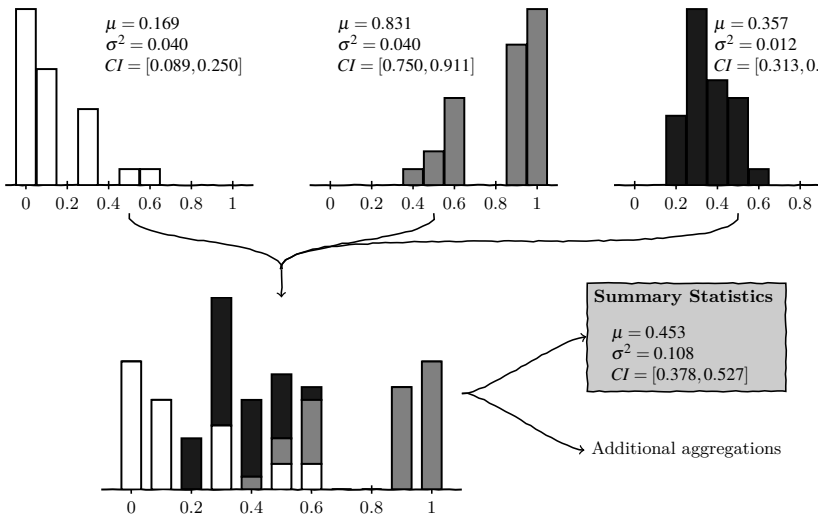


Figure 8: Illustration of Bootstrapped aggregation of three distributions. Simple summary statistics (mean, variance and 95% confidence interval presented for each distribution).

Bootstrapping is a simulation process that involves repeated sampling (in this case with replacement) of a sample set with the aim of generating a bootstrap sample from a distribution. This bootstrap sample can be used to estimate the underlying probability distribution function that generated the data as well as any other summary statistics. Importantly, bootstrapping provides a way to generate distributions that are proportional and thus un-weighted by the original sample sizes thereby facilitating un-weighted aggregation[3]. Bootstrapped distributions can be aggregated (added together) to yield accumulated child distributions that retain the combined properties of both parents (see Figure 8). As a stochastic process, repeated calculations will yield slightly different outcomes. Nevertheless, the more bootstrap samples are collected, the greater the bootstrap distributions will reflect the underlying Score distribution and provided the number of drawn samples is sufficiently large (e.g. 10,000 re-samples), repeated outcomes will converge.

To reiterate, the advantage of bootstrapping data before concatenating (or averaging) versus simply concatenating data from multiple sources together, is to ensure that source data are all of exactly the same sample size (so as to not weight more heavily towards the more populous source(s)[4]). Bootstrapping also provides a mechanism for propagating all distribution information throughout an aggregation hierarchy and ensures that estimates of variance derived from child distributions are on a consistent scale[5]. The latter point is absolutely critical if variance is going to be used to inform a Confidence Rating system and confidence intervals.

Minimum operator procedures are supported by filtering on the lowest performed indicator prior to bootstrapping. Importantly, the bootstrapping routine simply provides a mechanism to collate all sources together to yield a super distribution. Thereafter, the joint distribution can be summarized in what ever manner is deemed appropriate (arithmetic, geometric, harmonic means, medians, variance, range, quantiles etc). Moreover, different levels of the aggregation can be summarized with different statistics if appropriate.

[3] technically, all equally weighted rather than un-weighted
[4] Such weightings should be handled in other ways if at all
[5] Variance is inversely proportional to sample size

2.3.2 Weights

Standard bootstrapping yields equally weighted distributions, however, specific weighting schemes can also be easily applied by bootstrapping in proportion to the weights. For example, to weight one parent twice as high as another, simply collect twice as many re-samples from the first distribution. To ensure that all resulting distributions have the same size (by default 10,000 items), the number of bootstrap samples collected (n) from each of the (p) parent distributions (i), given the weights (w_i) is calculated as:

$$n_i = (S/p) \times w_i$$

where S is the target size (10,000) and . indicates the ceiling. Qualitative data (such as ratings) can also be incorporated by enumerating the categories before bootstrapping.

In addition to allowing expert driven weights that govern the contribution of different items during aggregations, it is possible to weight according to relative spatial areas during spatial aggregations. Currently, all Sites are equally weighted when aggregating to Zone level and all Zones equal when aggregating to Whole of Harbour level. That means that small Zones have an equal contribution as large Zones despite representing a smaller fraction of the water body. Area based weights could be applied such that Sites and Zones contribute in proportion to relative areas.

Weights are defined by a user editable configuration file that is similar in structure to the Water Quality guidelines file.

2.3.3 Expert interventions

The ability for experts and Report Card managers to intervene (exclude or overwrite) Scores/Grades at any Spatial/Measure scale is essential to maintain the quality of a Report Card in the event of unrepresentative or suspect data. The current system is able to support expert interventions in the form of exclusions and overwrites. For example, after reviewing the QAQC, an expert can elect to exclude one or more Measures (or Subindicators etc) from one or more spatial scales. Such interventions are specified via a user editable configuration files[6] (csv) that is similar in structure to the Water Quality guidelines file.

The essential component of this configuration file is that it allows a user to specify what Data are to be excluded or replaced. These can be at any of the levels of the Measure hierarchy (Measures, Sub-indications and Indicators) and any level of the Spatial hierarchy (Sites, Zones and Whole Bay). Settings pertaining to levels further along the aggregation hierarchies have precedence. For example, if Chemicals are excluded (or overridden) in a particular Zone, then all Chemical Measures within all Sites will be excluded irrespective of what the settings are for any specific Measure/Site.

2.3.4 Scores and Grades

The double hierarchy Bootstrap aggregation described above, yields **Score** distributions for each Measure-level/Spatial-level combination. The location and scale of each distribution can thus be described by its mean and variance. Mean **Scores** are then converted into a simple five-point alphanumeric **Grade** scale (and associated colors) using a control chart (see Table 4).

The control chart grade boundaries adopted for the current report (presented in Table 4) are not consistent with the Gladstone Healthy Harbour Partnership and Darwin Harbour Report Cards. Broadly, they represent two levels (Poor and Very Poor) under the Guideline values and three above (Satisfactory, Good and Very Good). The grade boundaries are usually determined by expert panel to ensure that the range of indices represented by each grade classification is congruent with community

[6] Since aggregation occurs across two hierarchies (the Measure hierarchy and the Spatial hierarchy - see Figures 6 and 7), two configuration files are necessary.

Table 4: GBR Report Card Grade scale control chart.

Score	Grade	Description
≥ 0.80	A	Very Good
$\geq 0.60, < 0.80$	B	Good
$\geq 0.40, < 0.60$	C	Satisfactory
$\geq 0.20, < 0.40$	D	Poor
< 0.20	E	Very Poor

interpretation of a letter grade report cards. It is far less clear how estimates of uncertainty can be incorporated into such a grading scheme in a manner that will be intuitive to non-technical audiences. That said, statistical uncertainty is just one of many sources of un- certainty that should be captured into a confidence or certainty rating. Hence any expectations of presenting uncertainty in a quantitative manner may well be unrealistic anyway.

2.3.5 Certainty rating

Incorporating an estimate of scale (variance) into a certainty or confidence rating necessitates re-scaling the estimates into a standard scale. In particular, whereas a scale parameter of high magnitude indicates lower degrees of certainty, for a certainty rating to be useful for end users, larger numbers should probably represent higher degrees of certainty. Thus, the scaling process should also reverse the scale. Furthermore, variance is dependent on the magnitude of the values.

In order to re-scale a scale estimate into a certainty rating, it is necessary to establish the range of values possible for the scale estimate. Whilst the minimum is simple enough (it will typically be 0), determining the maximum is a little more challenging depending on the aggregation algorithm (bootstrapping, Bayesian Network etc). One of the advantages in utilizing proportional distributions (such as is the case for a Bayesian Network or a re-sampled bootstrap distribution) is that the scale parameter for the single worst case scenario can be devised (once the worst case scenario has been determined) independent of sample sizes or weightings. In most situations this is going to be when the distribution comprises equal mass at (and only at) each of the two extremes (for example, values of just 0 and 1).

The measure of confidence rating discussed above is purely an objective metric derived from the variance in the aggregation hierarchy. It is completely naive to issues such as missing data, outliers and Limit of Detection issues - the influences of which on a confidence rating are necessarily subjective. A full Confidence Rating would combine these objective variance component with additional subjective considerations such as climatic and disturbance information, and the perceived influence of missing, Limit of Detection and outlying data. Hence, the statistical scaled statistical variance would form just one component in the Confidence Rating system.

The bootstrap aggregation method provides a mechanism for estimating variance from which to build such an expert considered Confidence Rating system.

2.3.6 Confidence intervals

Confidence intervals (CI) represent the intervals in which we have a certain degree of confidence (e.g. 95%) that repeated estimates will fall. Hence the 95% CI of the mean is the range defined by the quantiles representing 95% of repeated estimates of the mean.

To calculate 95% confidence intervals for bootstrap aggregated distributions (e.g. Site 1/Chemical distribution), we repeatedly[7] draw a single sample from each of the constituent distributions (e.g. a single value from the Site 1 Ammonia, Chlorophyll-a and NOx distributions) and from each set of draws,

[7]The more repeated draws the closer the distribution of means will converge. For the current project, the number of repeated draws is 10,000.

calculate the weighted [8] mean of the values. The 95% CI is thus calculated as the quantiles (p=0.025 and p=0.975) of the means.

[8]Weights according to the weights defined for that level of the aggregation hierarchy

The aggregation schedule can be summarized as:

A. Calculation of Zone level Score and Grades

1. Collect raw data (= **Measures**) at each fixed monitoring site and compare individual observations to associated guideline/benchmark/reference or set of expectation ranges
2. Create indexed data as an expression of degree of difference (*scaled modified amplitude method*) to yield a **Score** for each **Measure** per sampling location (e.g. Site) (applies to *Measures* in all *Indicators*, Water Quality). In the absence of guidelines (e.g. Measures within Plankton), observed data are rescaled to a range defined by historical quantiles (20th and 80th percentiles) for each Measure.
3. Apply any expert opinion interventions
4. Combine **Measure** Scores into **Site**-level **Sub-indicator** Scores by averaging taking into account any weightings, i.e. aggregate into observation-level Sub-indicator Scores. This step involves **Bootstrapping** each input to distributions of 10,000 re-samples (or fewer if weighted), combining distributions and finally Bootstrapping again into a single 10,000 size distribution.
5. Combine **Sub-indicator** Scores into **Site**-level **Indicator** Scores by averaging, i.e. aggregate into Site-level Indicator Scores.
6. Convert Scores into coloured **Grades** (A-E) for visual presentation in report card

B. Calculation of Zone level Grades

1. Aggregate **Site**-level *Indicator* Scores from step A.5 into **Zone**-level *Indicator* Scores by averaging (incorporating spatial weights)
2. Aggregate **Zone**-level *Indicator* Scores into **Zone**-level *Component* Scores by averaging (incorporating weights)

C. Calculation of Whole Bay Grades

1. Aggregate **Zone**-level *Indicator* Scores from step B.1 into **Whole Bay**-level *Indicator* Scores by averaging (incorporating spatial weights)
2. Aggregate **Whole Bay**-level *Indicator* Scores into **Whole Bay**-level *Component* Scores by averaging (incorporating weights)

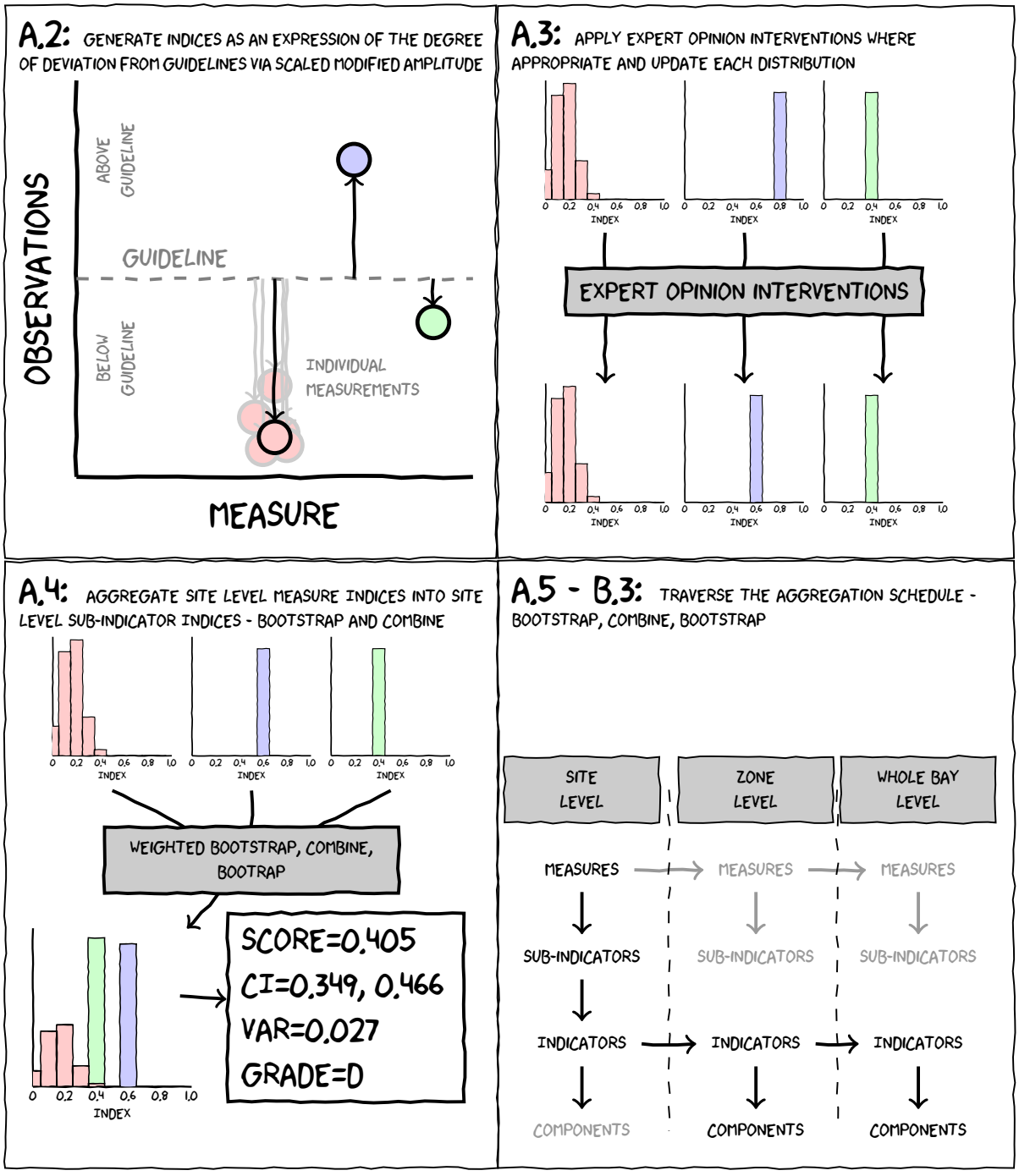


Figure 9: Schematic illustrating the major steps of the GBR Report Card. In this fabricated example, there are three Measures (Red, Green and Blue). Each of the Blue and Green Measures are represented by a single discrete observation, whereas the Red Measure is represented by a large collection of observations. Expert option intervened to lower the blue Measure distribution from observed values at 0.8 to 0.6.

2.3.7 Annual thresholds

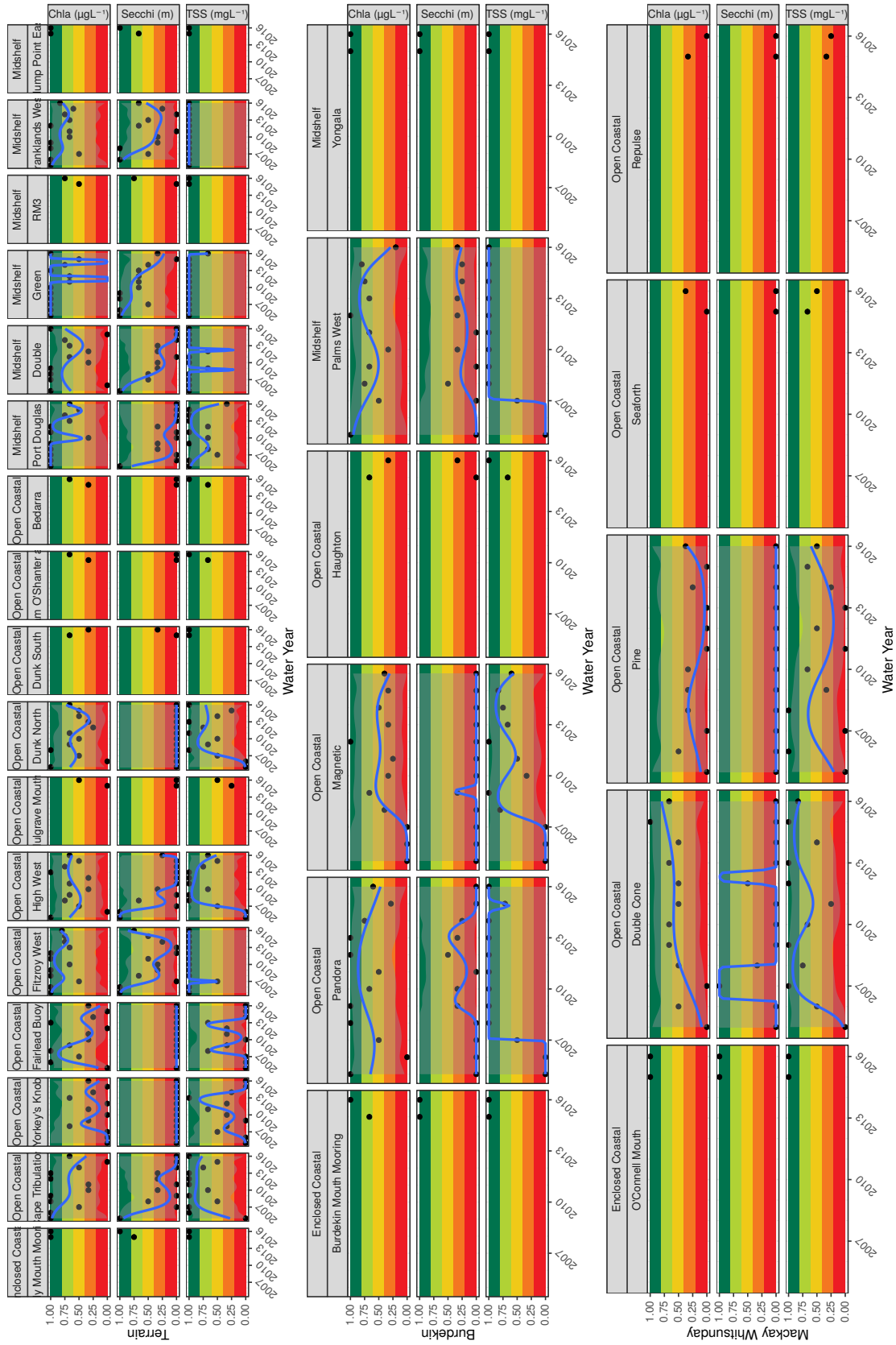


Figure 10: Time series (annual averages) of Binary indices of AIMS in situ samples. GAM smoothers applied.

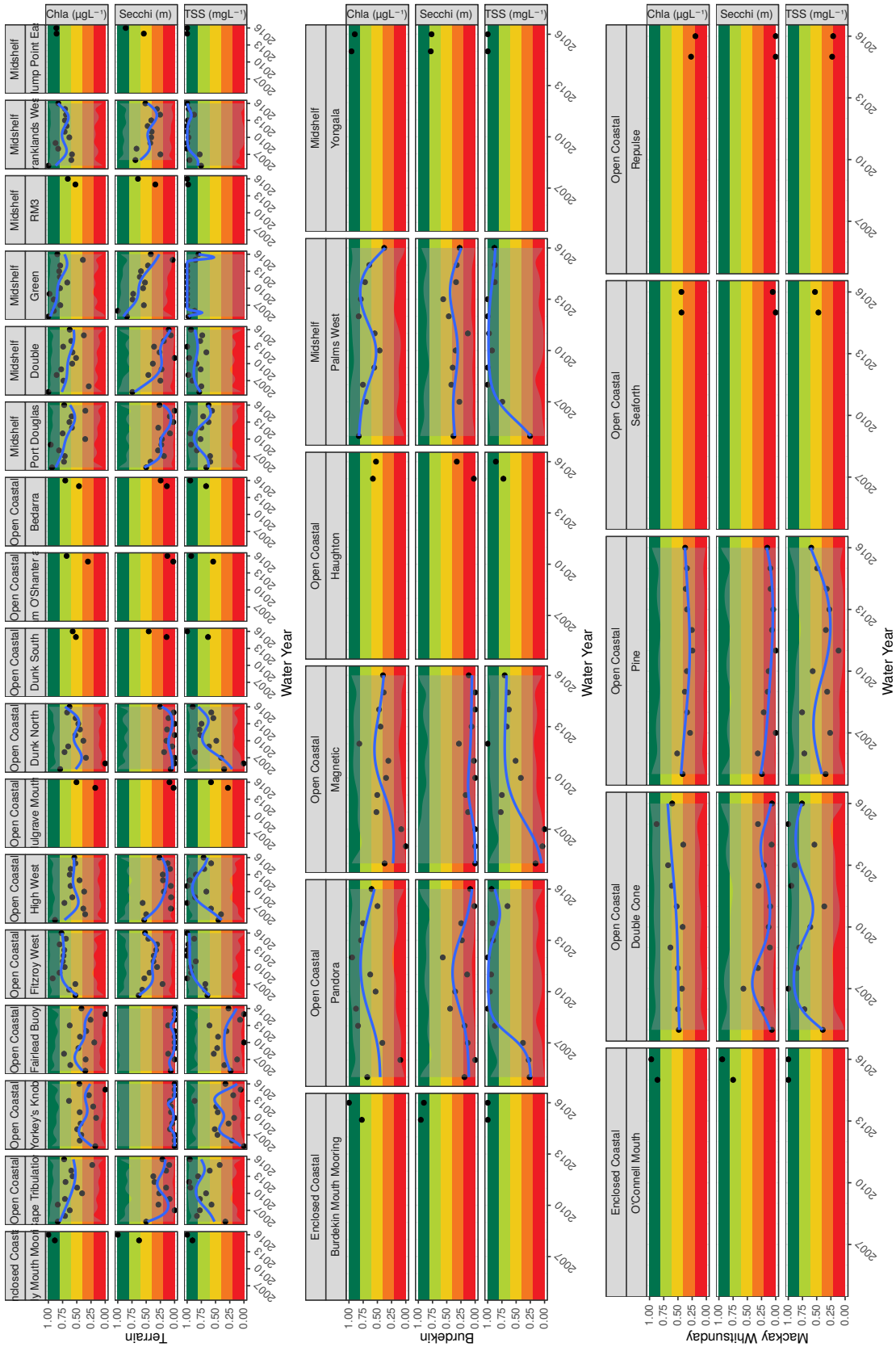


Figure 11: Time series (annual averages) of fsMAMP indices of AIMS in situ samples. GAM smoothers applied.

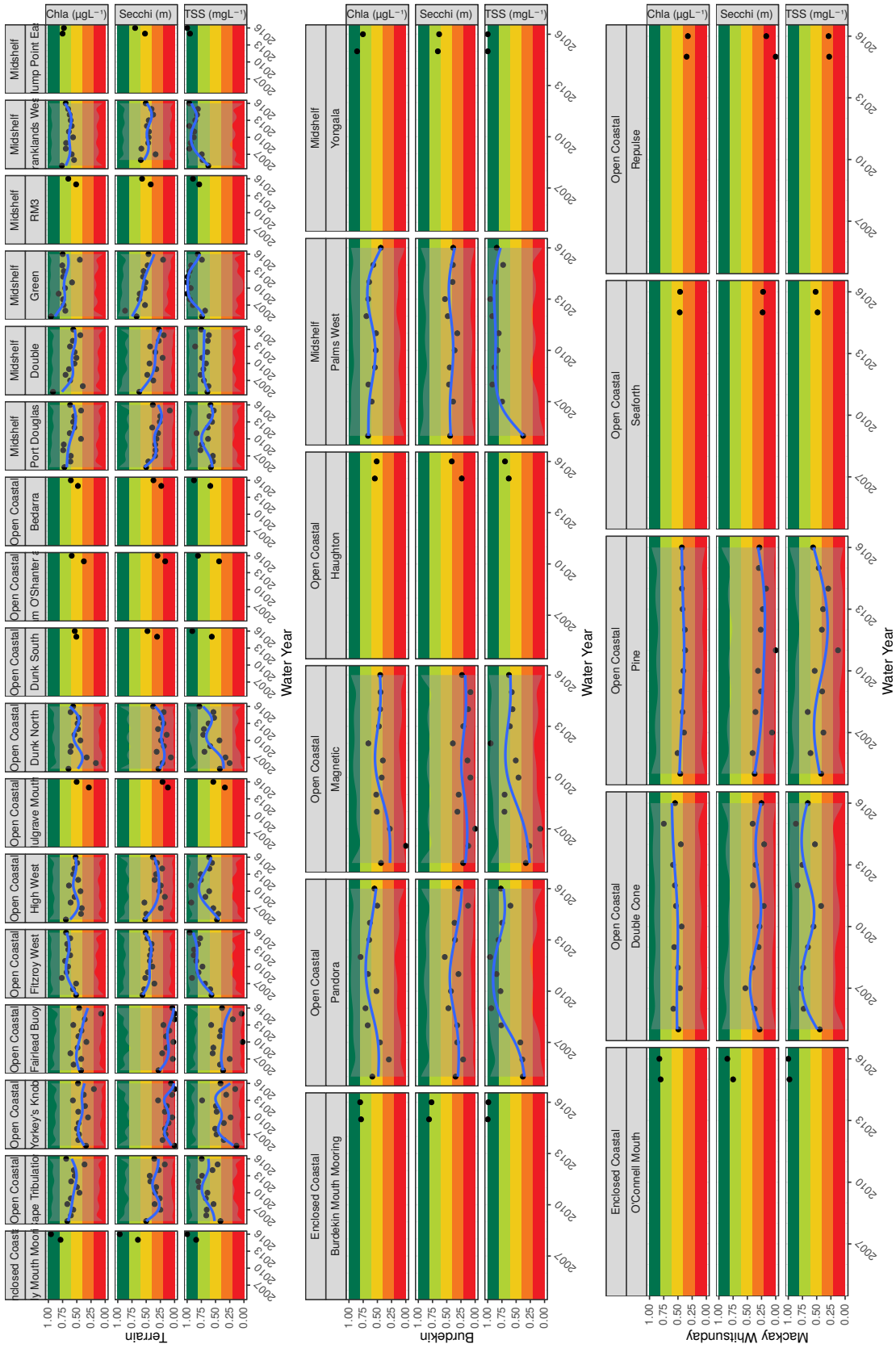


Figure 12: Time series (annual averages) of fsMAMP4 indices of AIMS in situ samples. GAM smoothers applied.

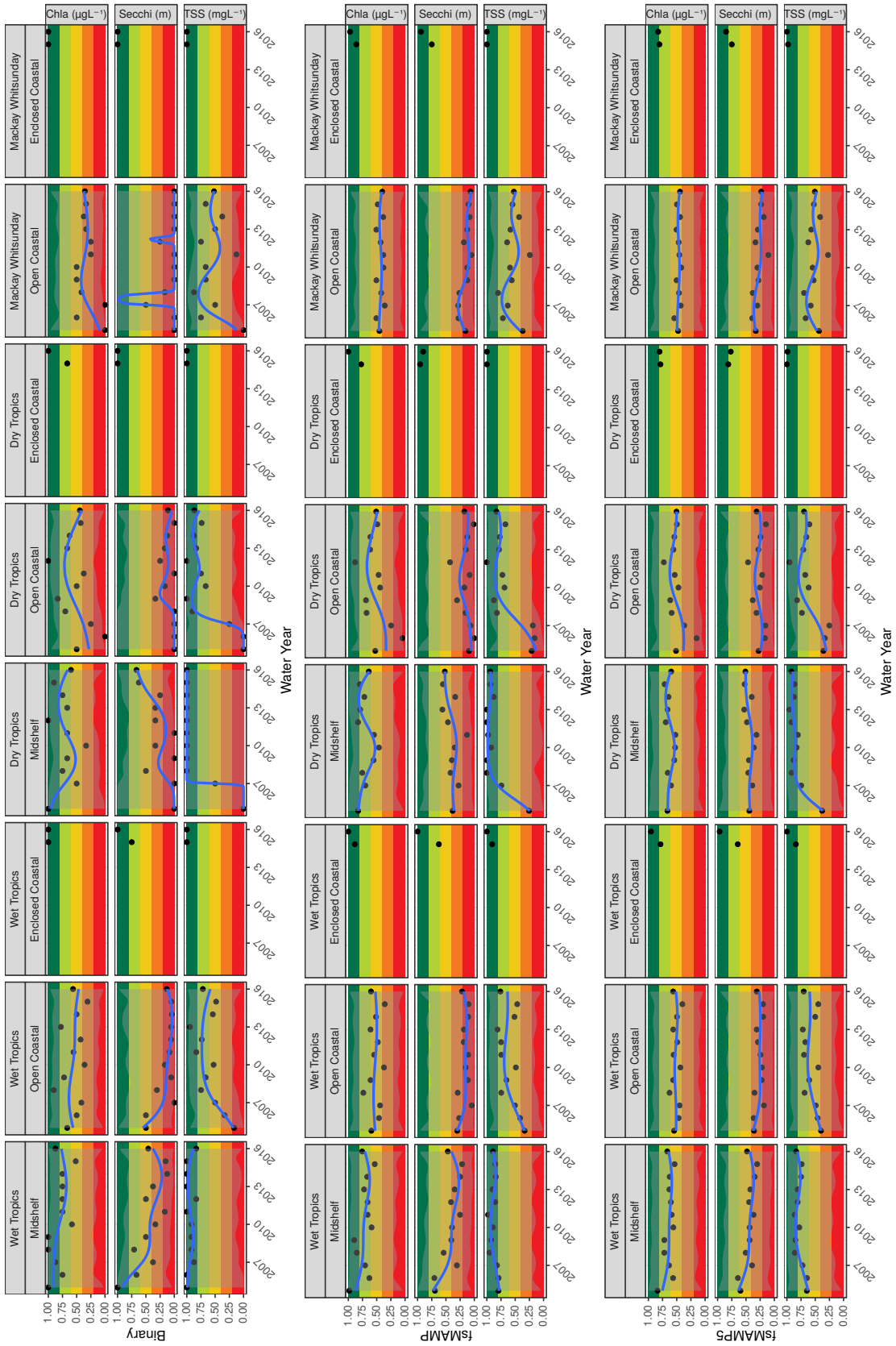


Figure 13: Time series (annual averages for each Zone) of various indices of AIMS in situ samples. GAM smoothers applied.

Maps

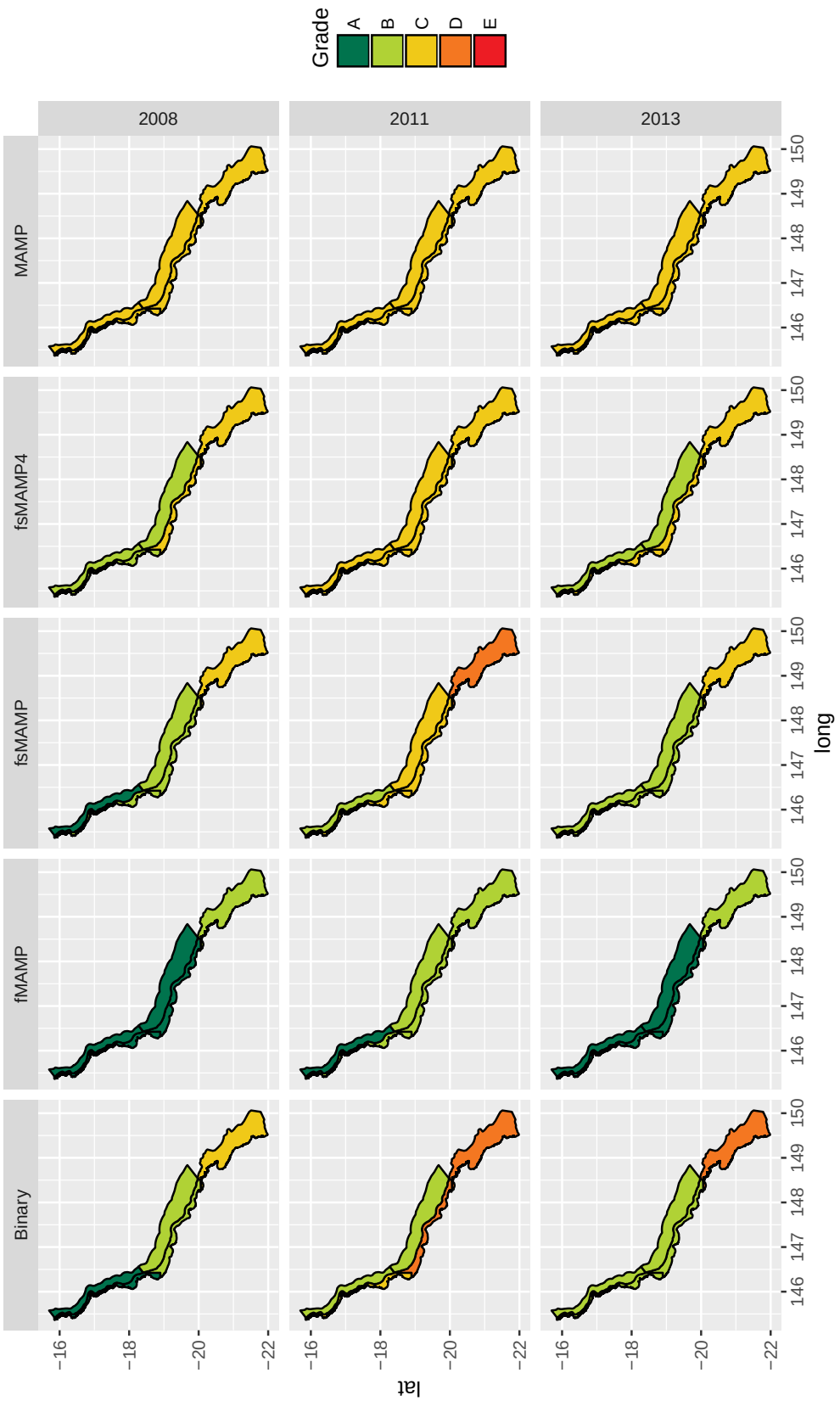


Figure 14: Chlorophyll index grades by zone for selected years.

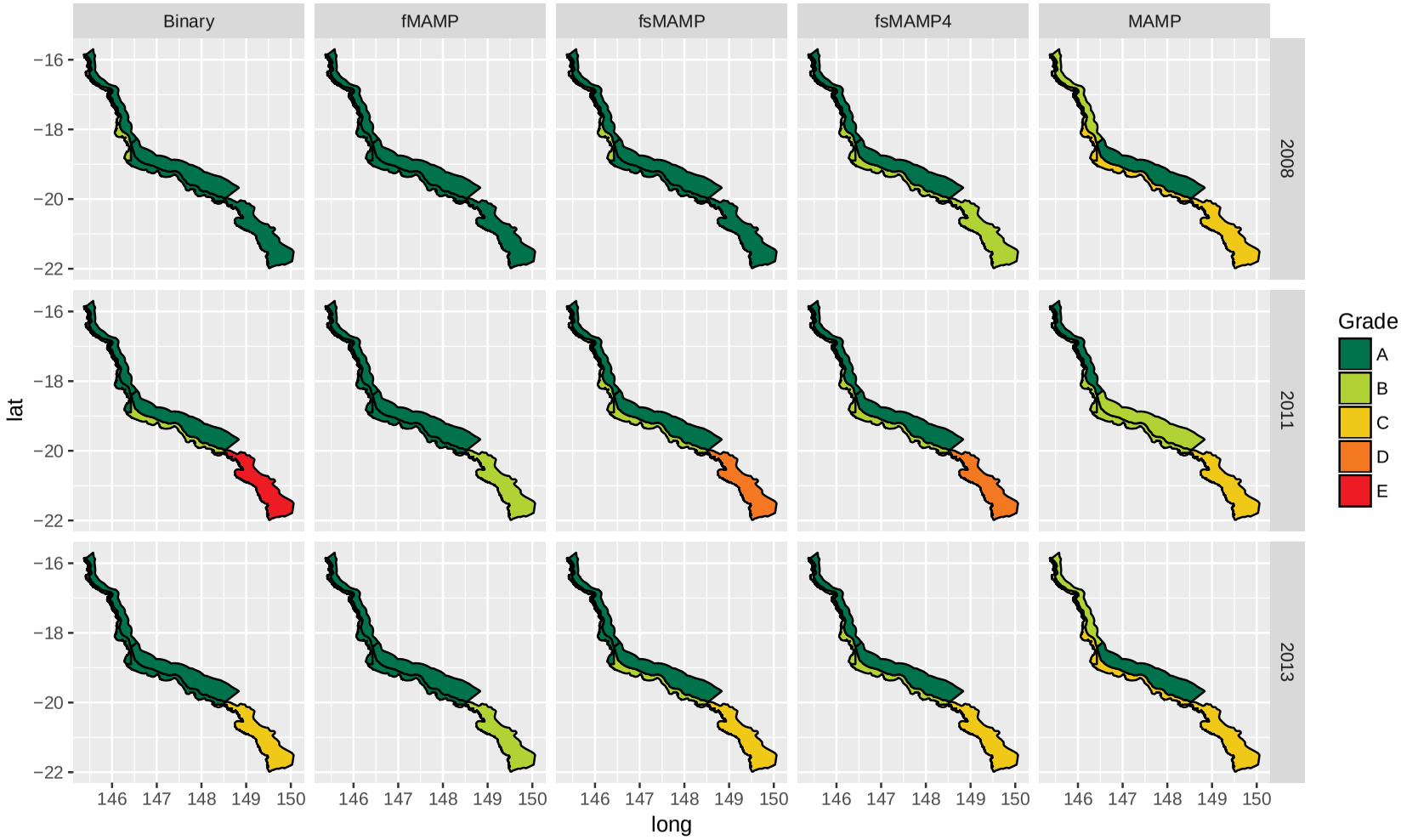


Figure 15: Suspended solids (Nap) index grades by zone for selected years.

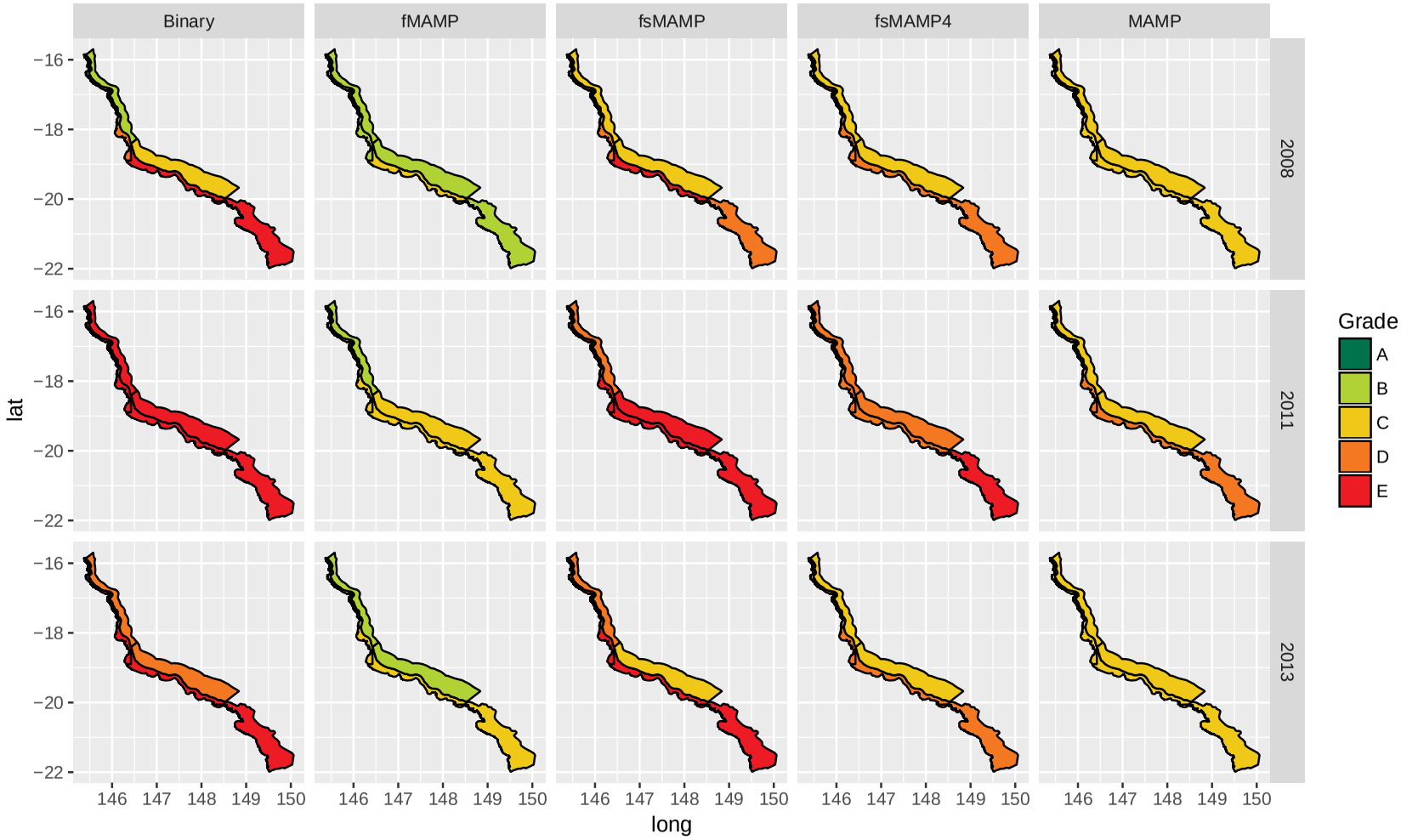


Figure 16: Secchi Depth (SD) index grades by zone for selected years.

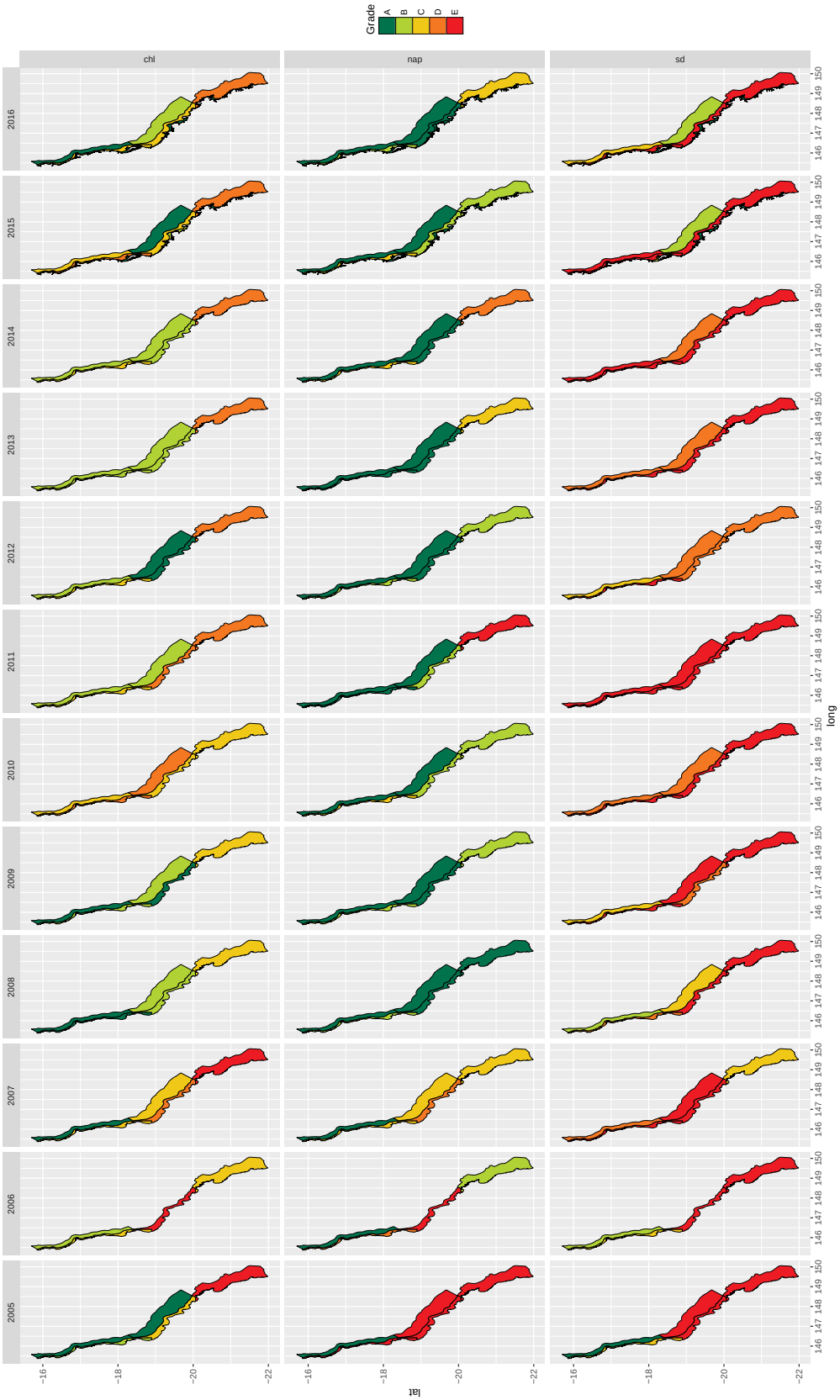


Figure 17: Binary index grades by zone.

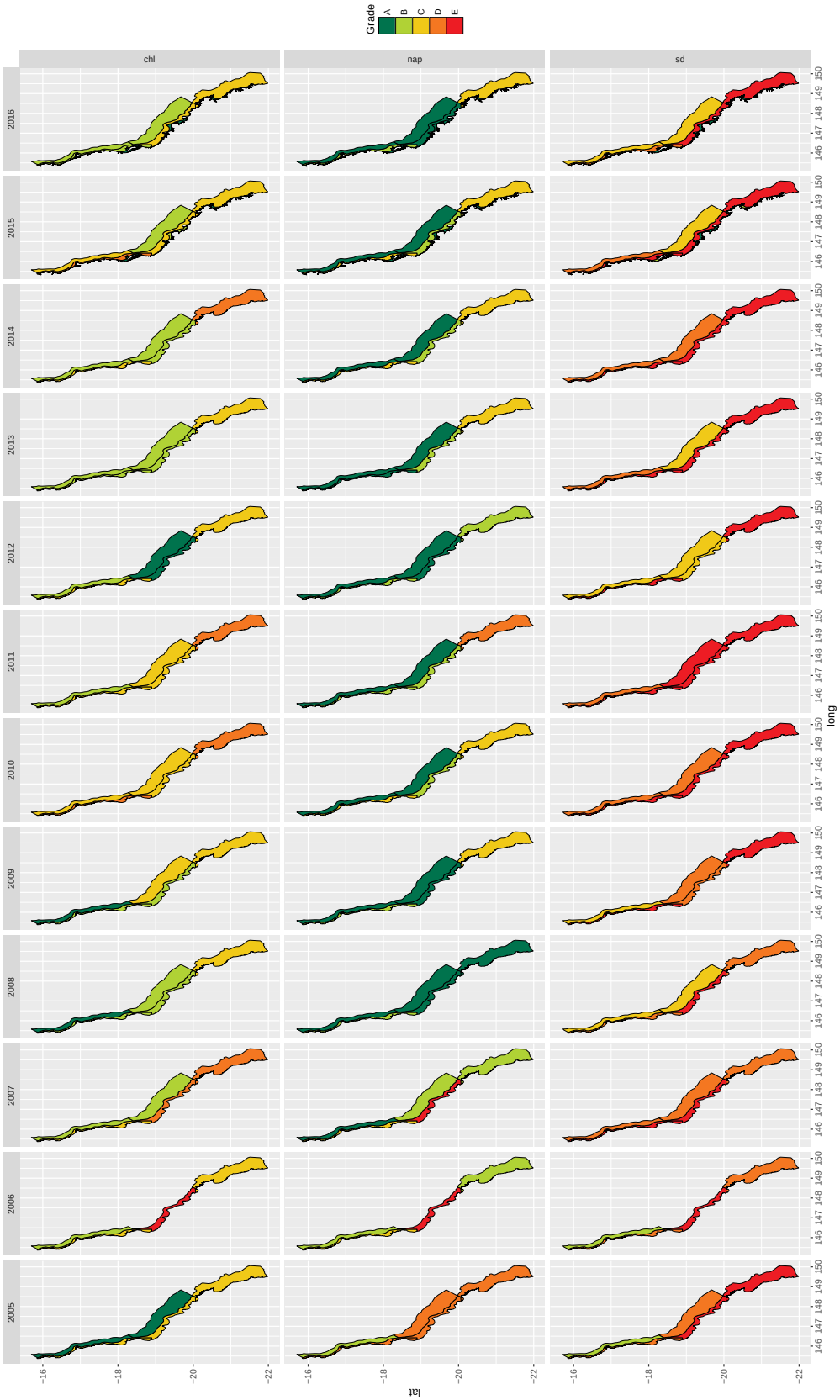


Figure 18: fsMAMP index grades by zone.

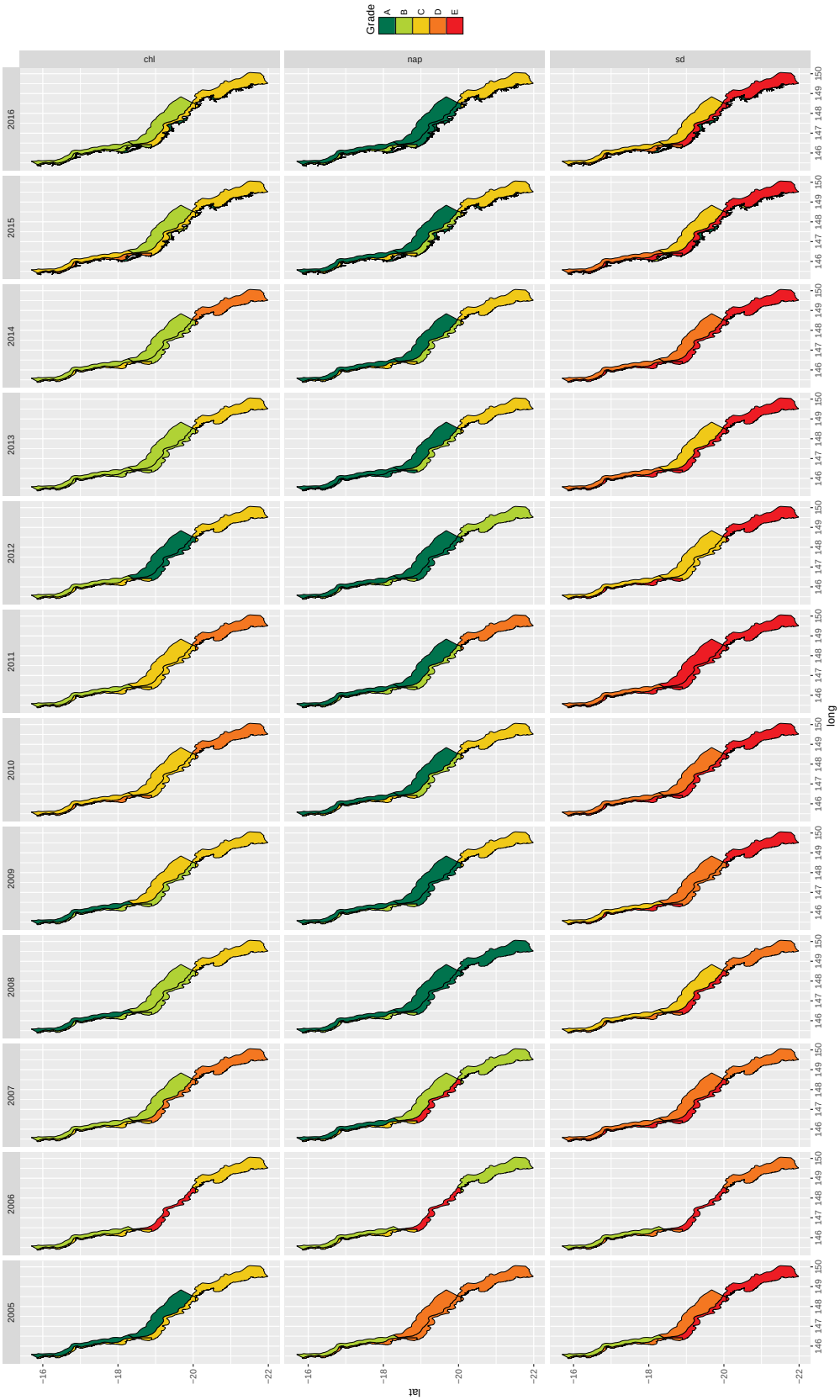


Figure 19: fsMAMP4 index grades by zone.

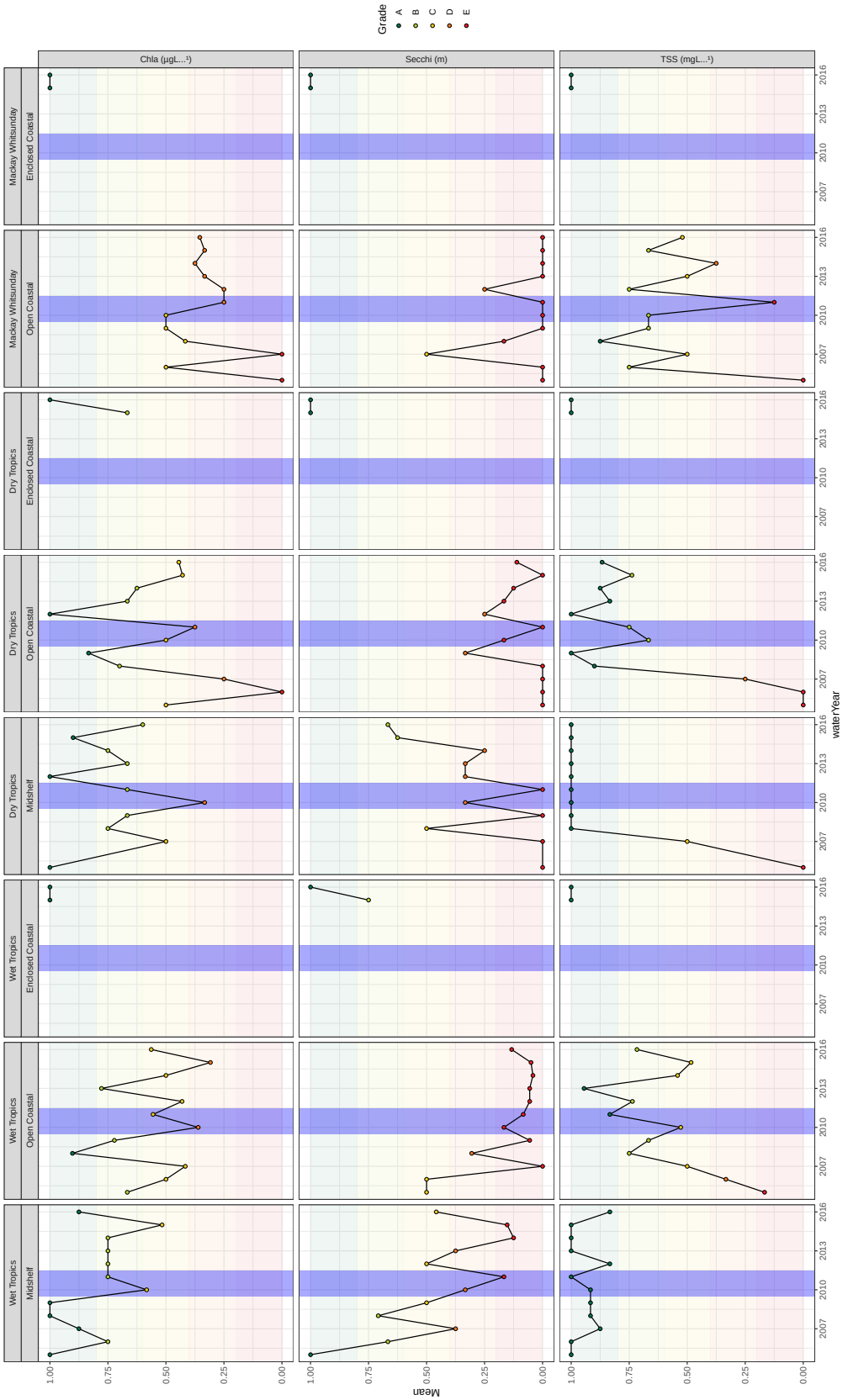


Figure 20: Time series of Binary index scores by zone. The blue vertical bar spans from mid 2009 to mid 2011.

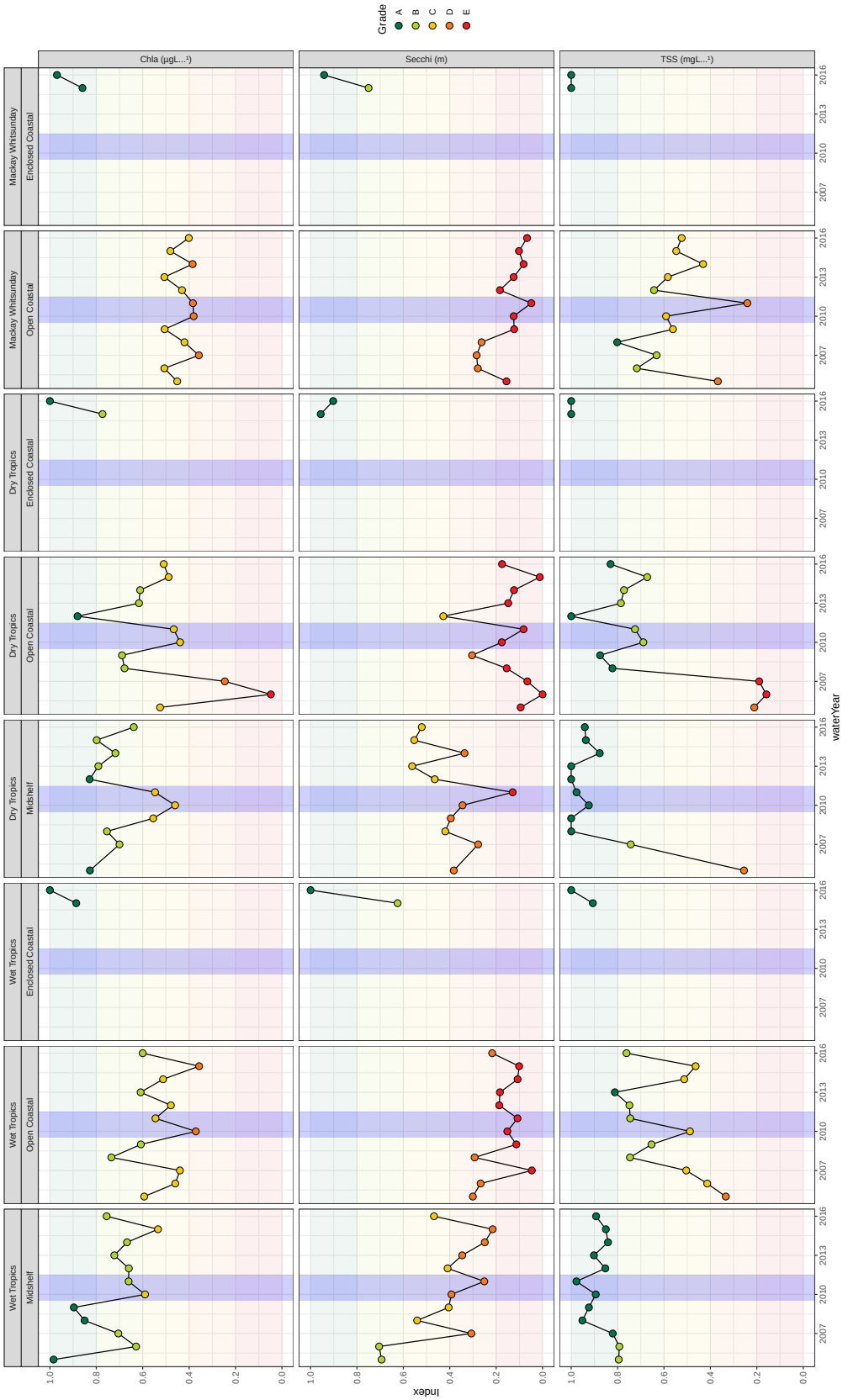


Figure 21: Time series of fsMAMP index scores by zone. The blue vertical bar spans from mid 2009 to mid 2011.

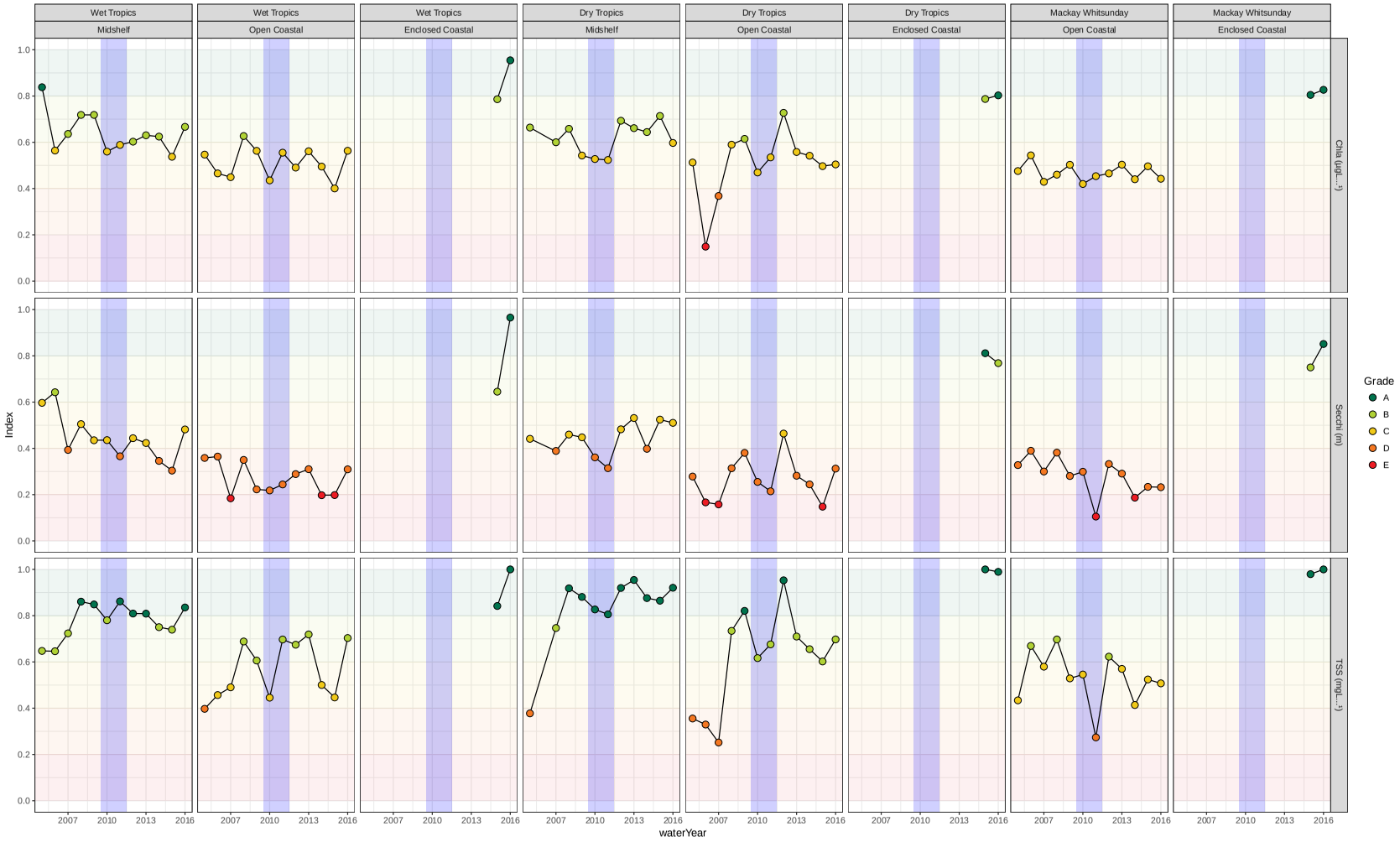


Figure 22: Time series of fsMAMP4 index scores by zone. The blue vertical bar spans from mid 2009 to mid 2011.

Bootstrapping

Measure hierarchy last (Figure 7, Option b)

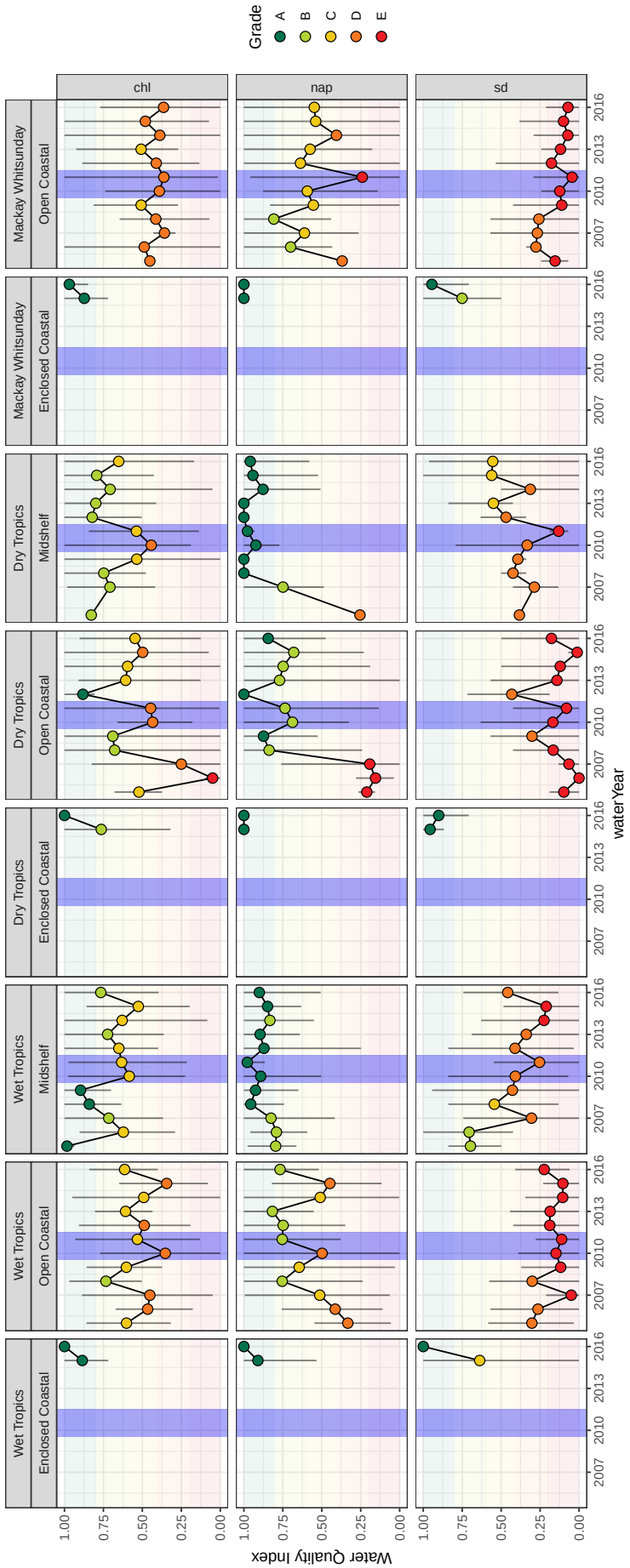


Figure 23: Time series of fsMAMP Measure index scores by zone. The blue vertical bar spans from mid 2009 to mid 2011.

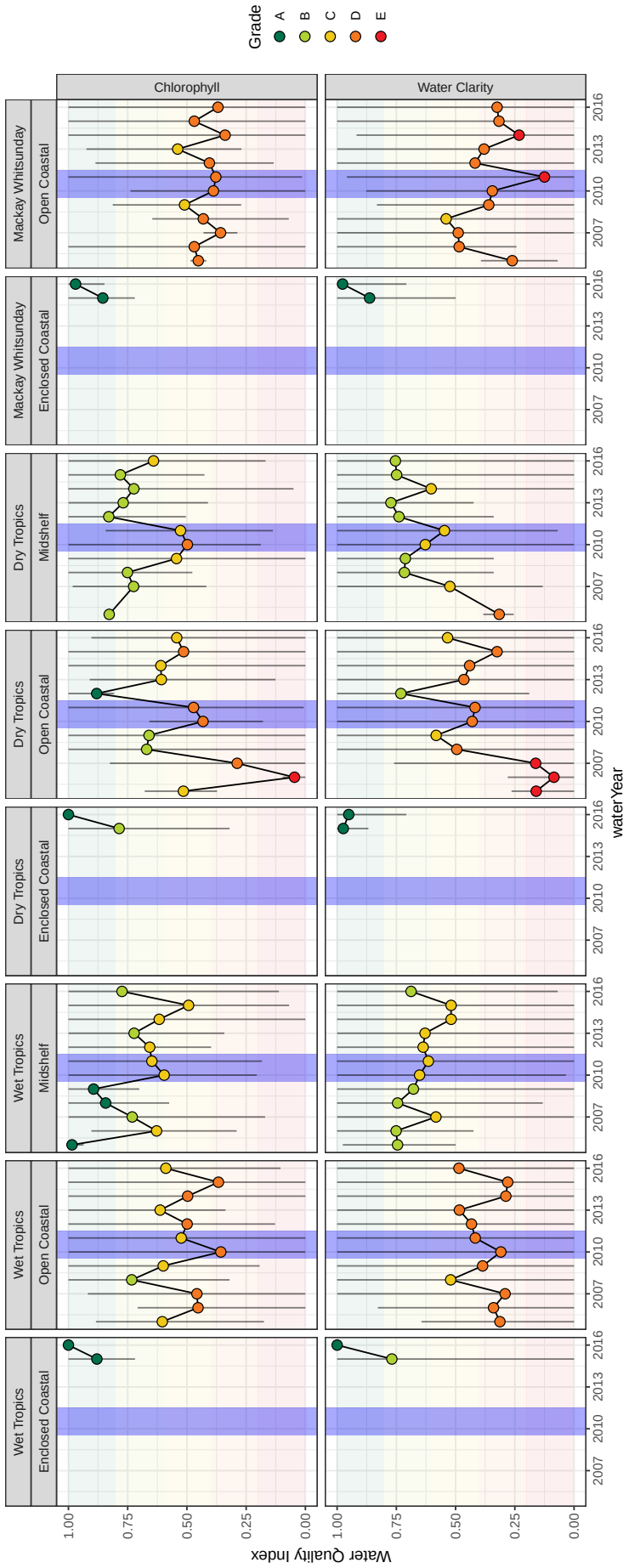


Figure 24: Time series of fsMAMP Subindicator index scores by zone. The blue vertical bar spans from mid 2009 to mid 2011.

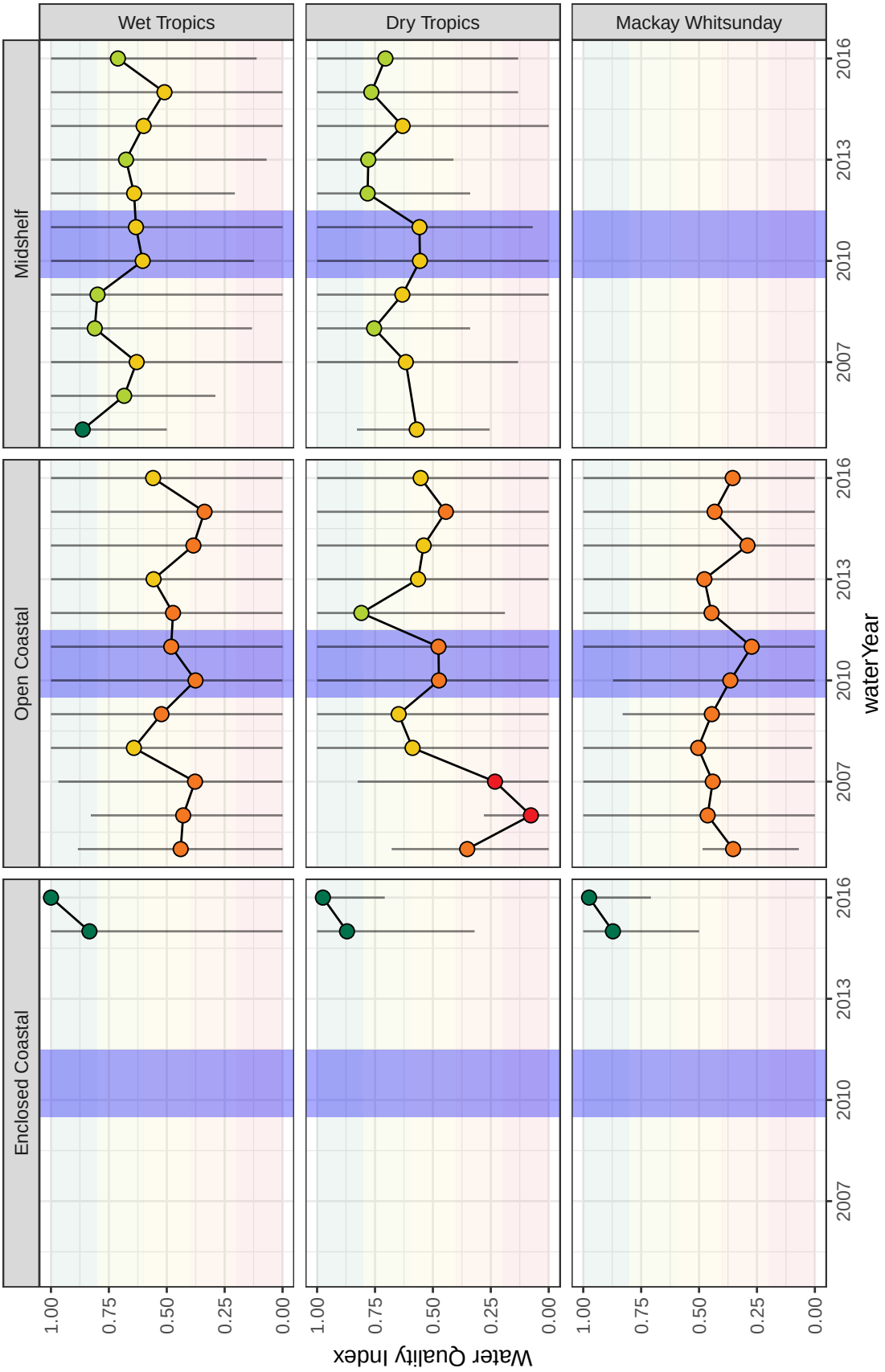


Figure 25: Time series of fsMAMP Indicator index scores by zone. The blue vertical bar spans from mid 2009 to mid 2011.

Spatial hierarchy last (Figure 7, Option a)

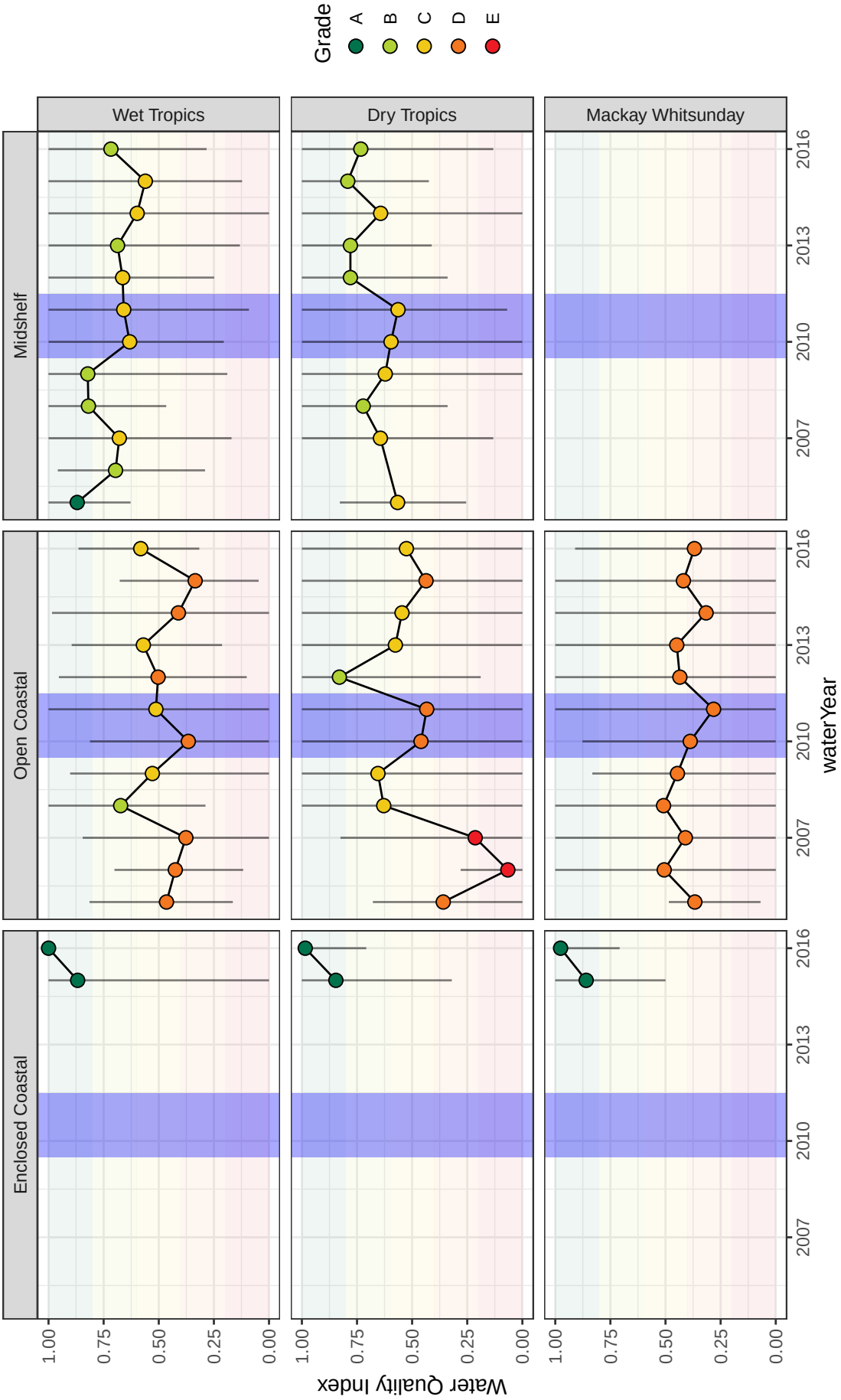


Figure 26: Time series of fsMAMP Indicator index scores by zone. The blue vertical bar spans from mid 2009 to mid 2011.

2.3.8 Seasonal thresholds

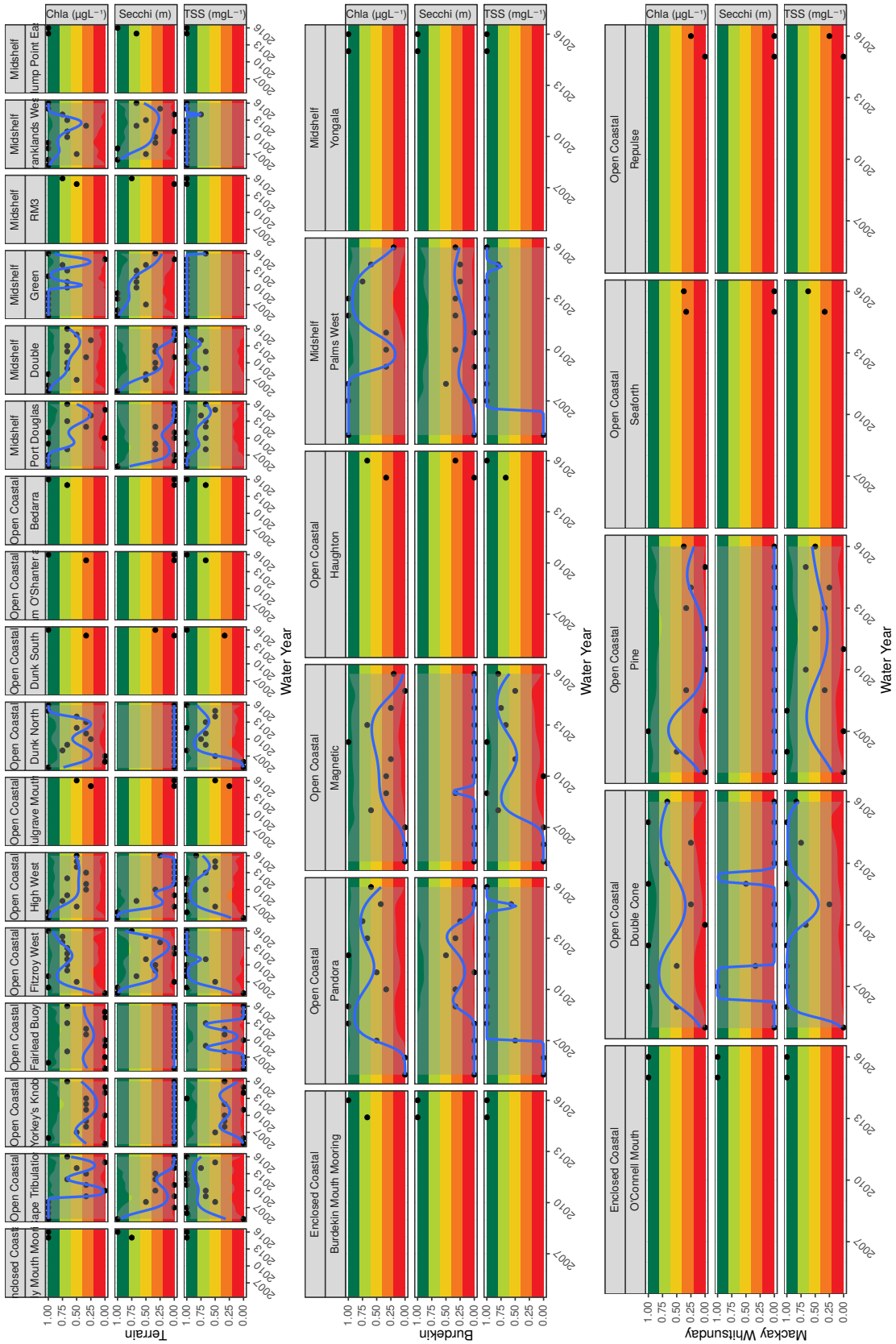


Figure 27: Time series (annual averages) of Binary indices (Seasonal thresholds) of AIMS in situ samples. GAM smoothers applied.

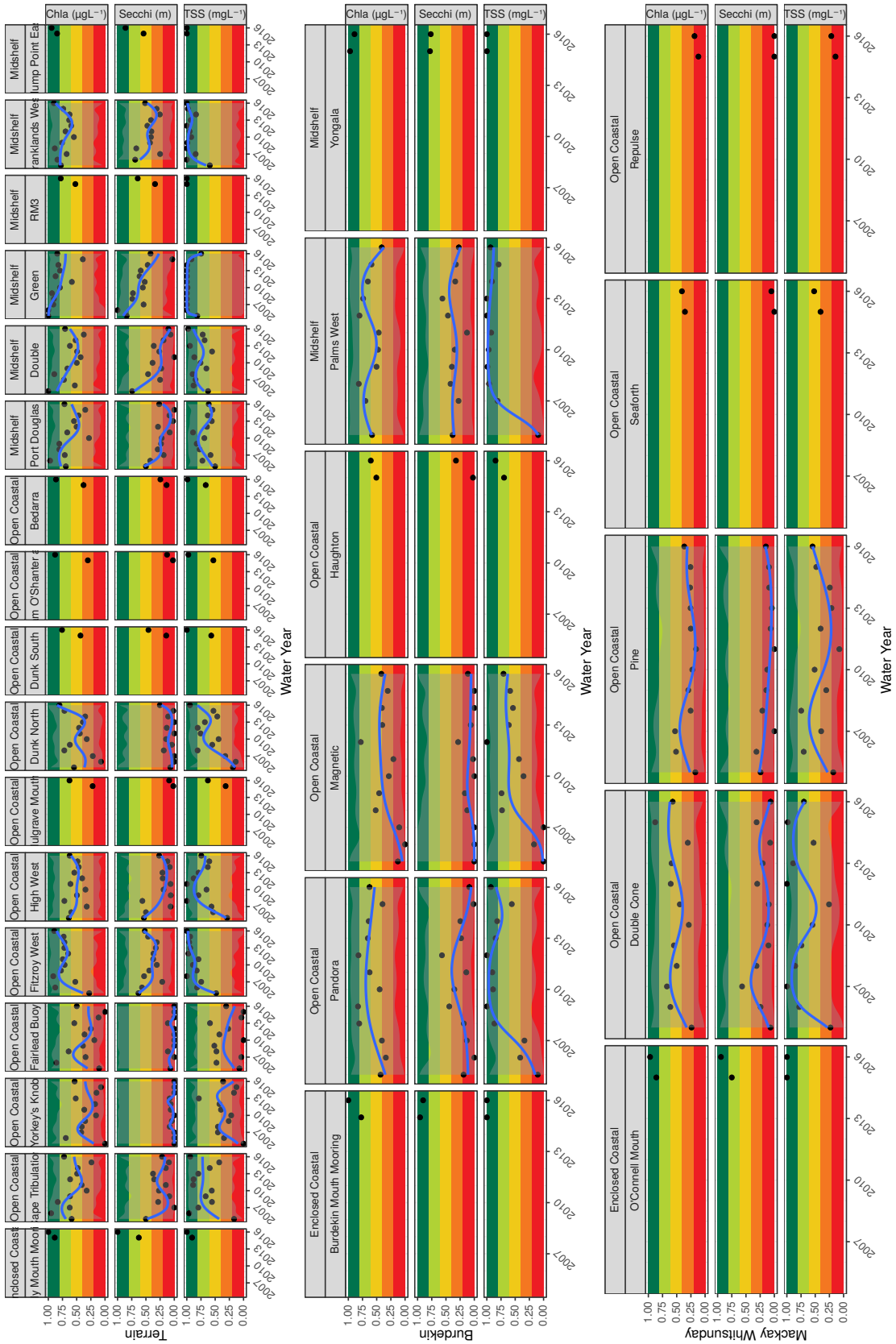


Figure 28: Time series (annual averages) of fsMAMP indices (Seasonal thresholds) of AIMS in situ samples. GAM smoothers applied.

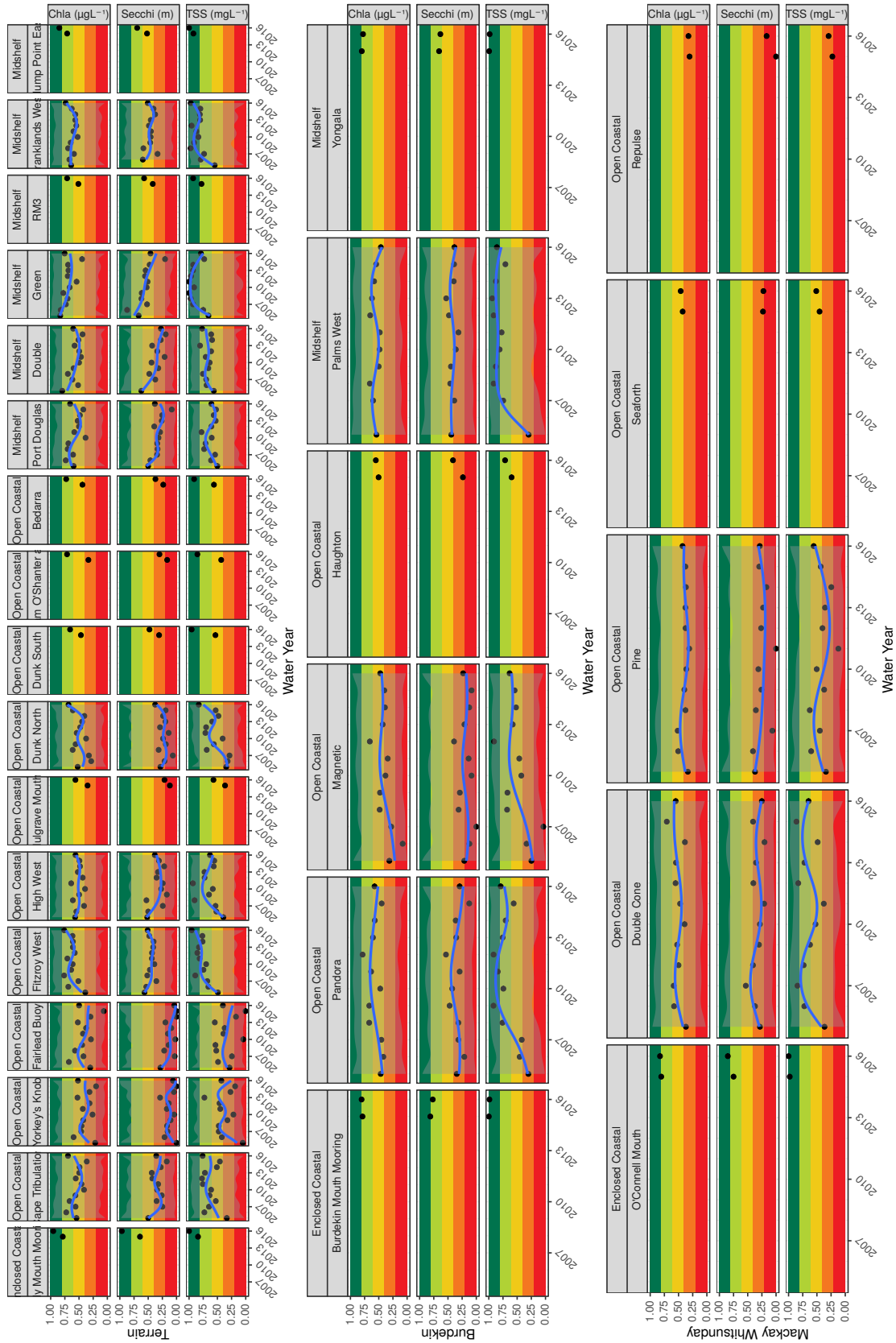


Figure 29: Time series (annual averages) of fsMAMP4 indices (Seasonal thresholds) of AIMS in situ samples. GAM smoothers applied.

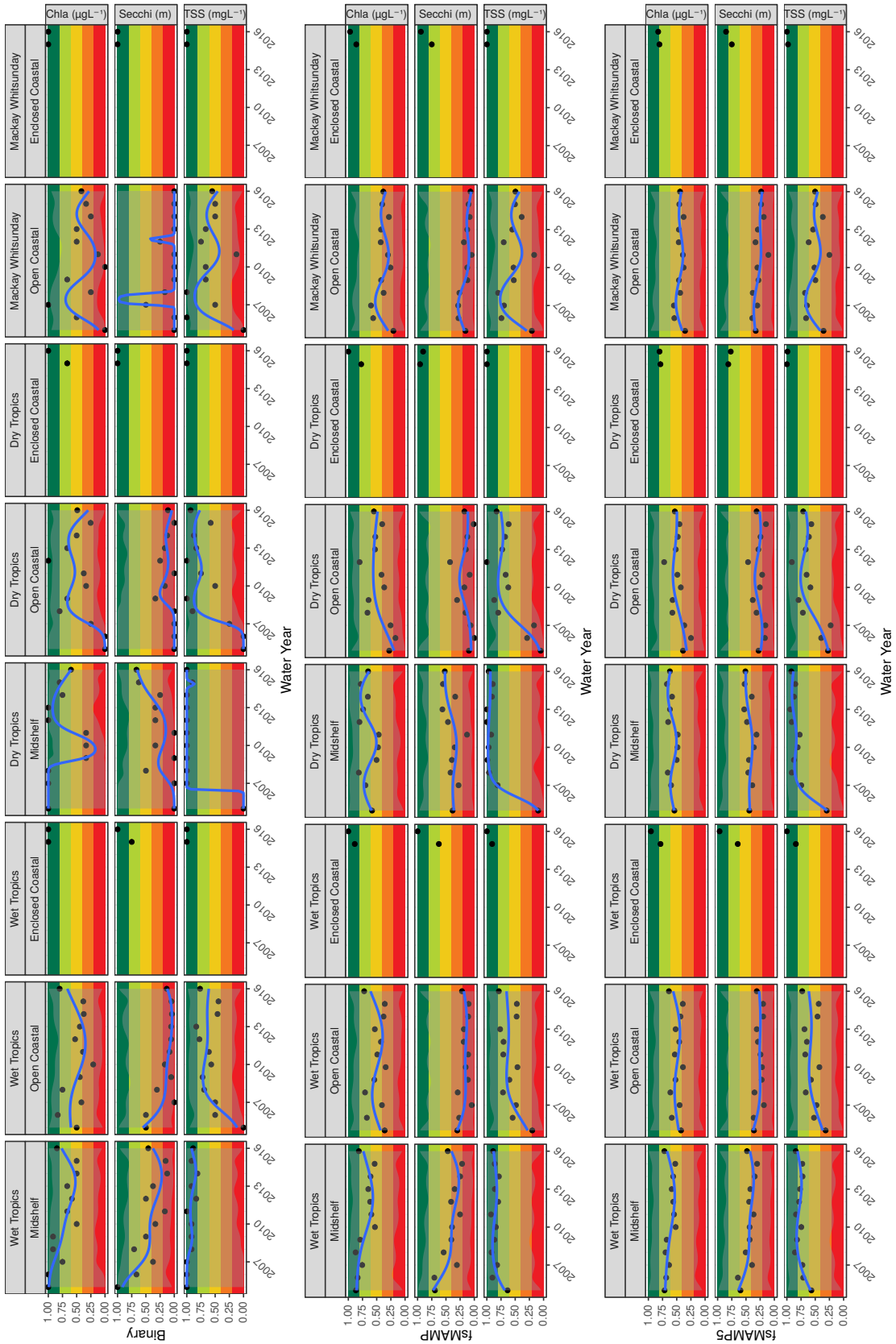


Figure 30: Time series (annual averages for each Zone) of various indices (Seasonal thresholds) of AIMS in situ samples. GAM smoothers applied.

Maps

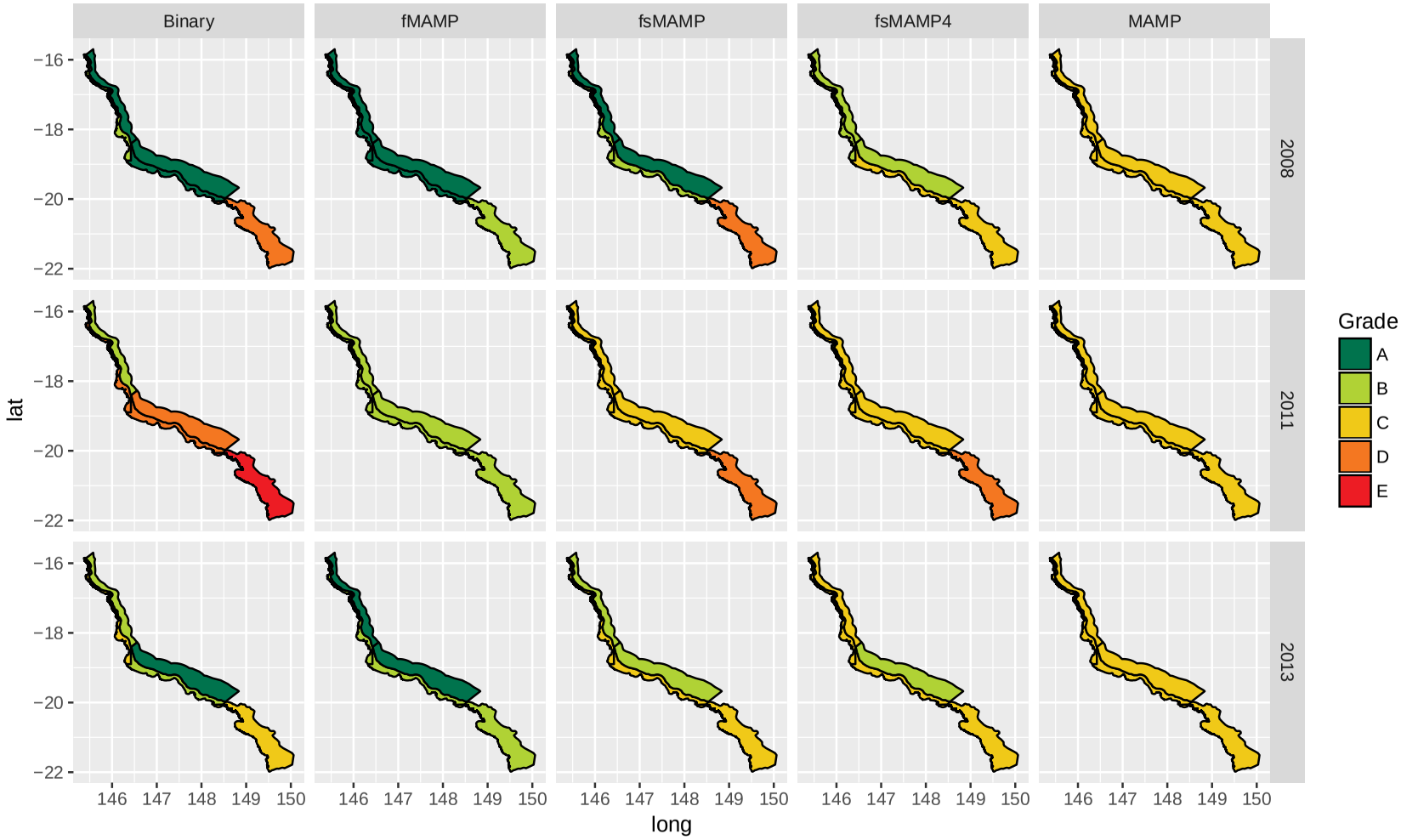


Figure 31: Chlorophyll index (Seasonal thresholds) grades by zone for selected years.

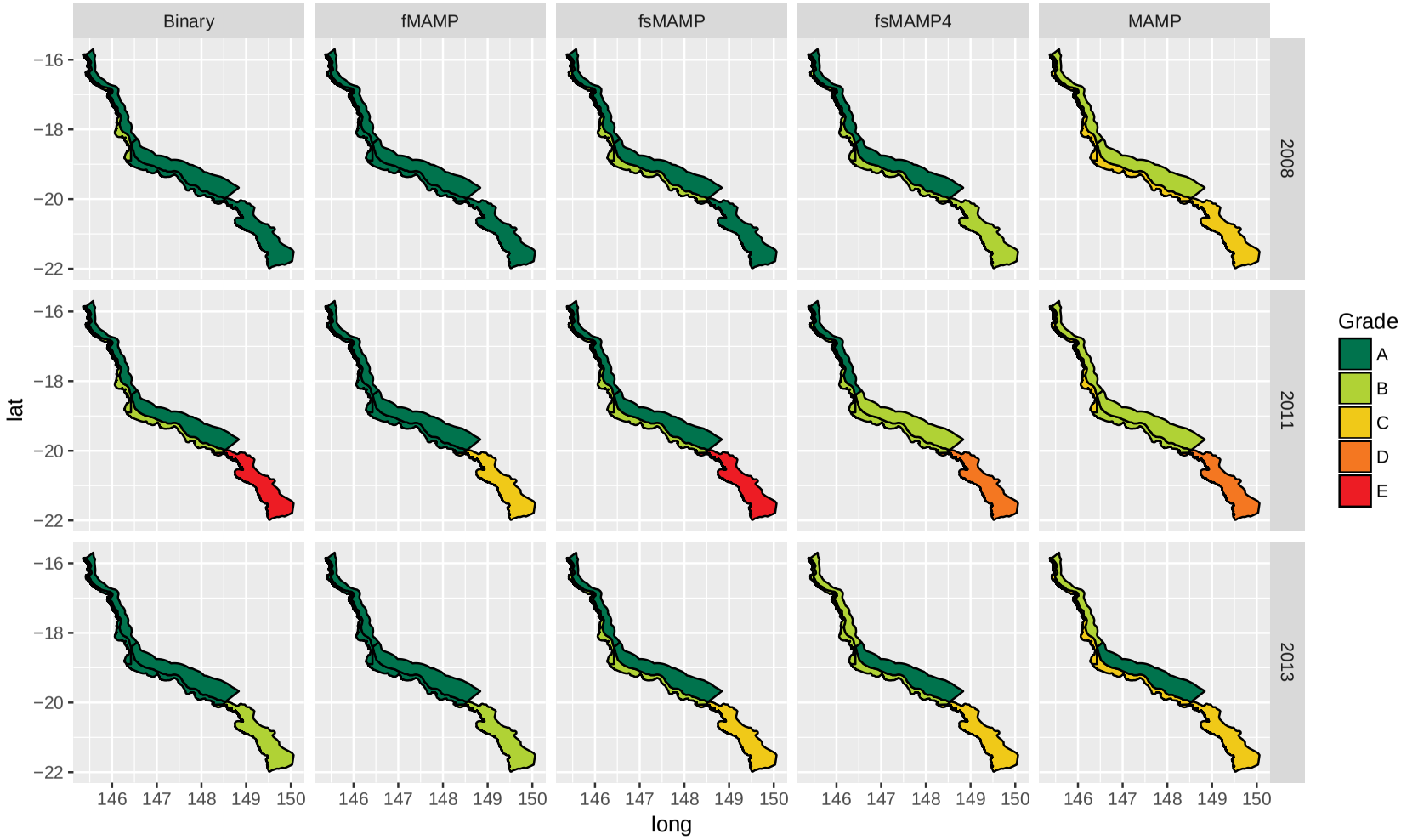


Figure 32: Suspended solids (Nap) index (Seasonal thresholds) grades by zone for selected years.

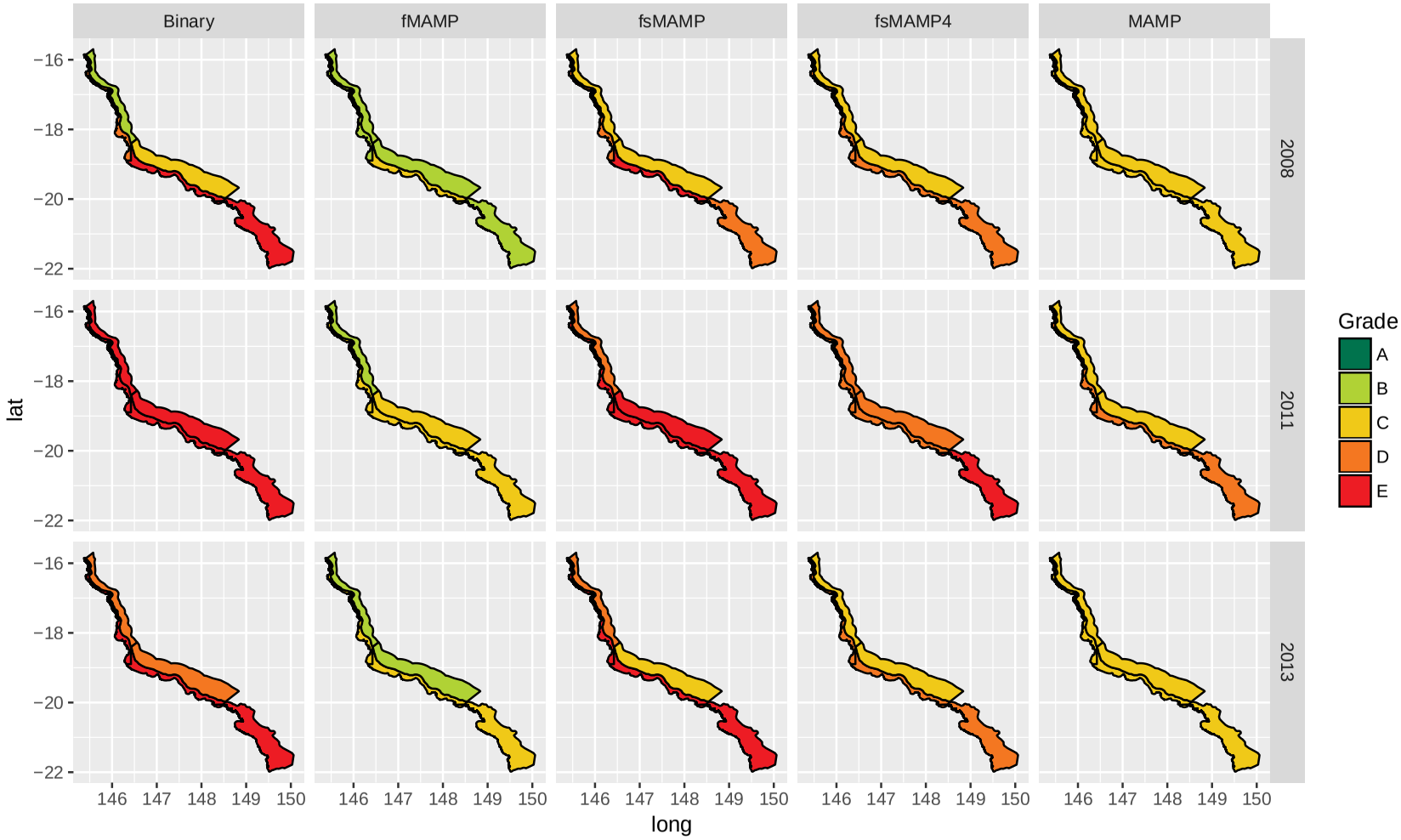


Figure 33: Secchi depth (SD) index (Seasonal thresholds) grades by zone for selected years.

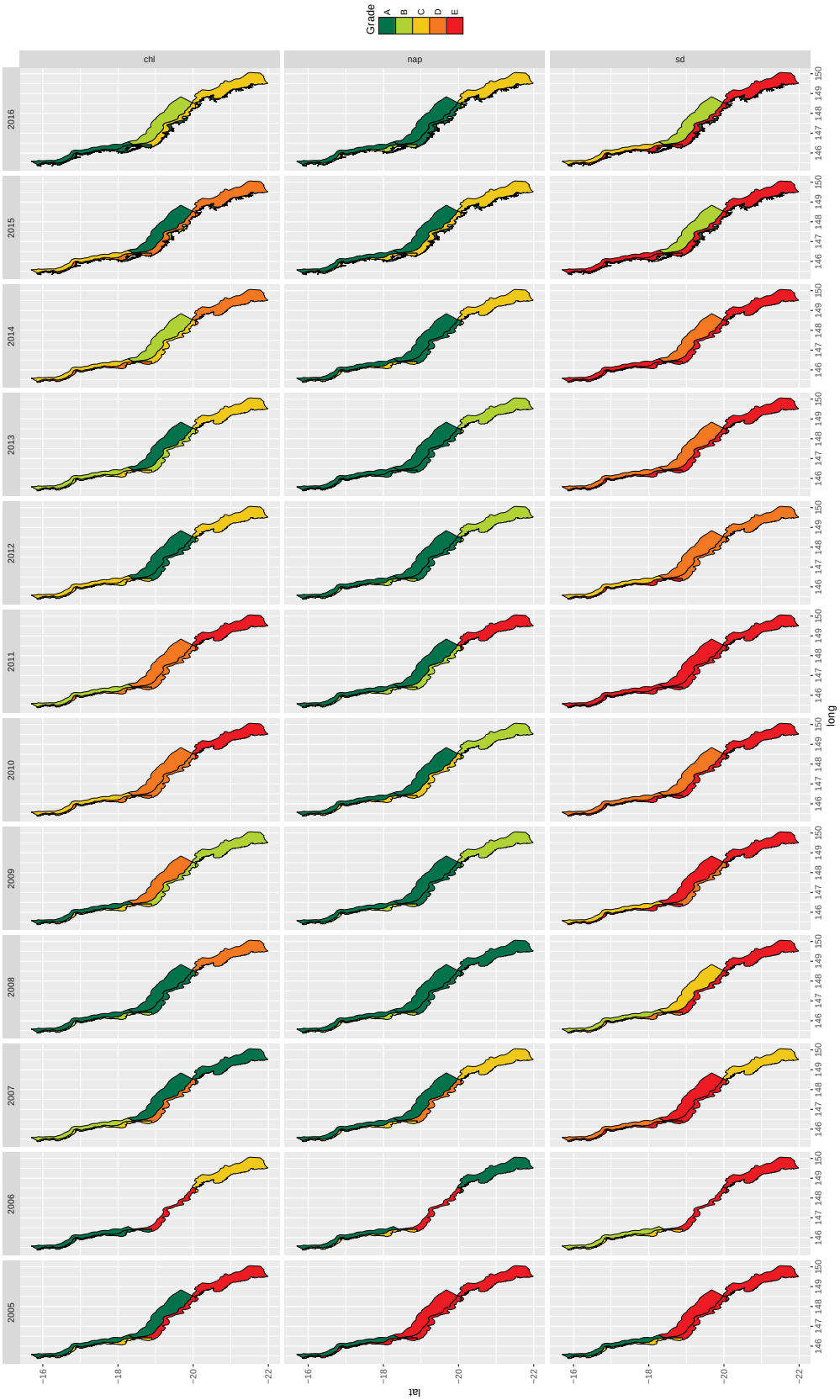


Figure 34: Binary index (Seasonal thresholds) grades by zone.

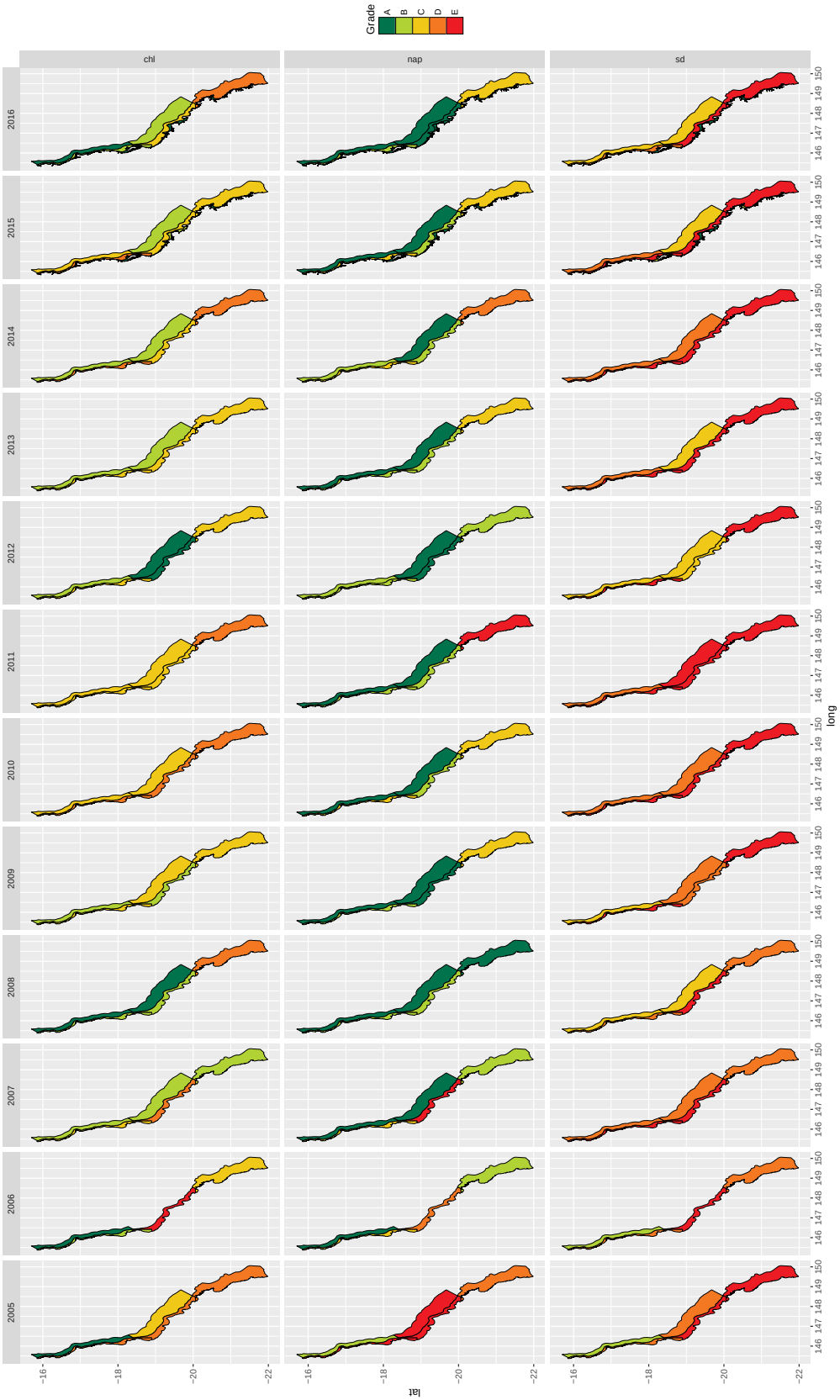


Figure 35: fsMAMP index (Seasonal thresholds) grades by zone.

Figure 36: fsMAMP4 index (Seasonal thresholds) grades by zone.

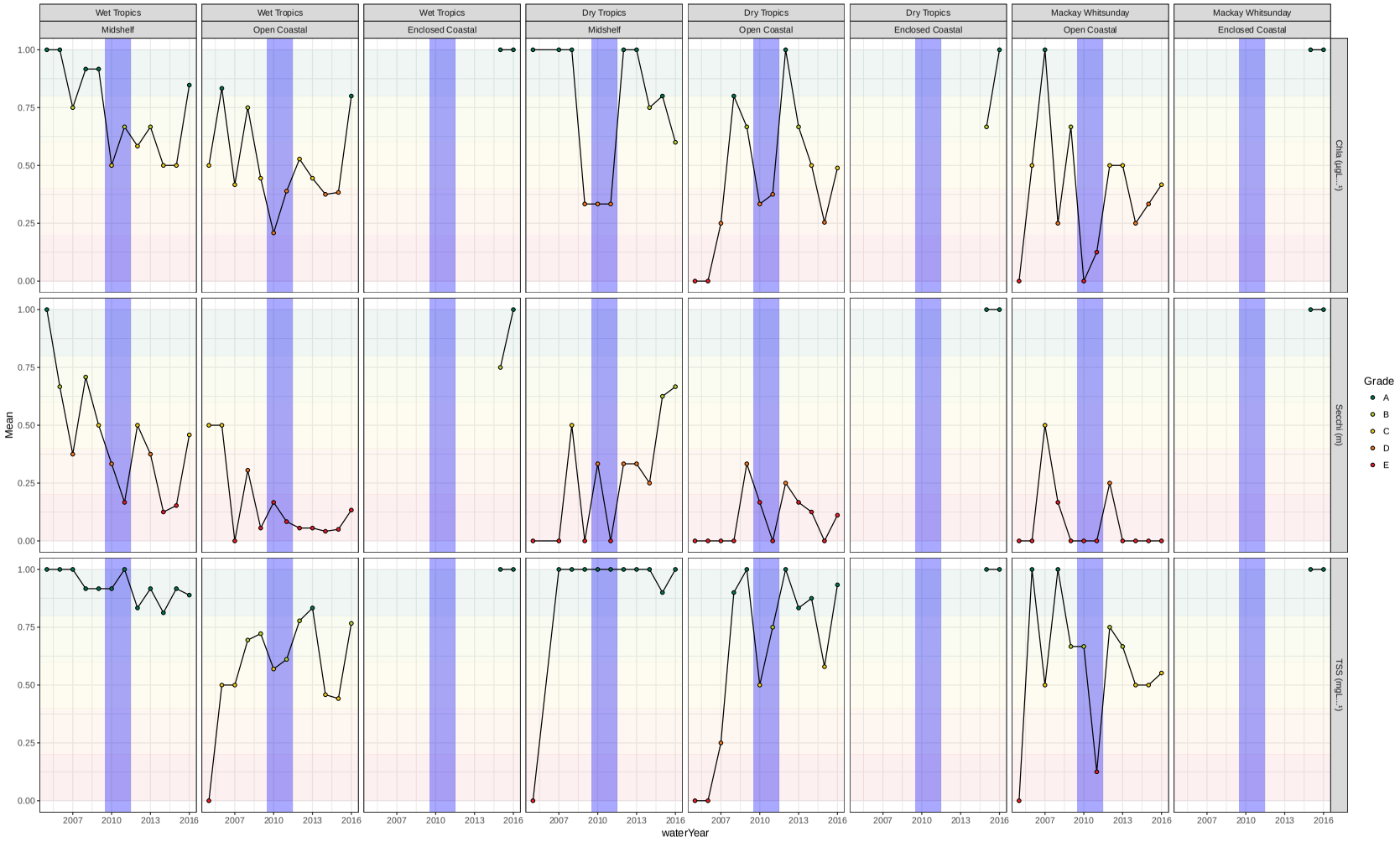


Figure 37: Time series of Binary index (Seasonal thresholds) scores by zone. The blue vertical bar spans from mid 2009 to mid 2011.

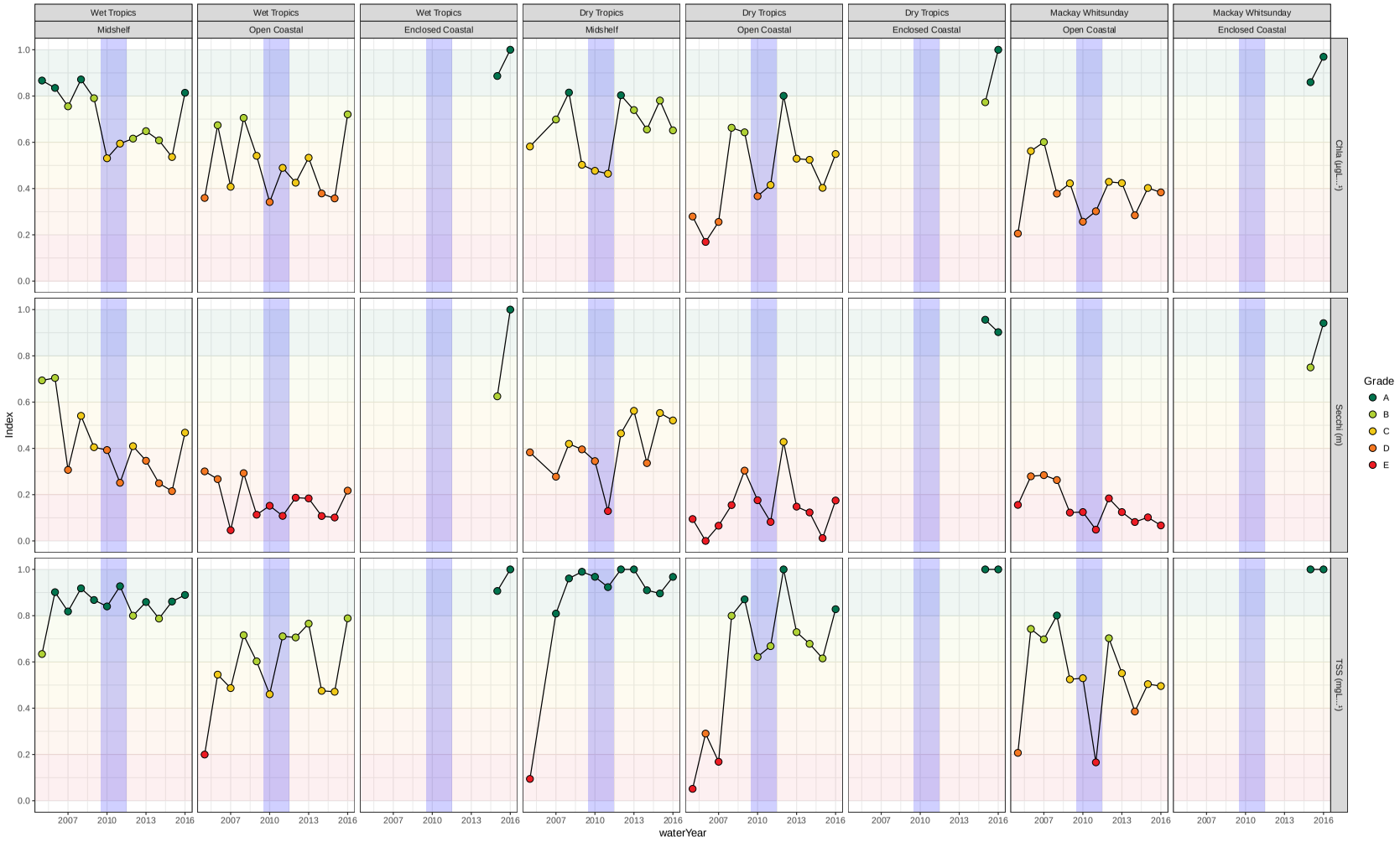


Figure 38: Time series of fsMAMP index (Seasonal thresholds) scores by zone. The blue vertical bar spans from mid 2009 to mid 2011.

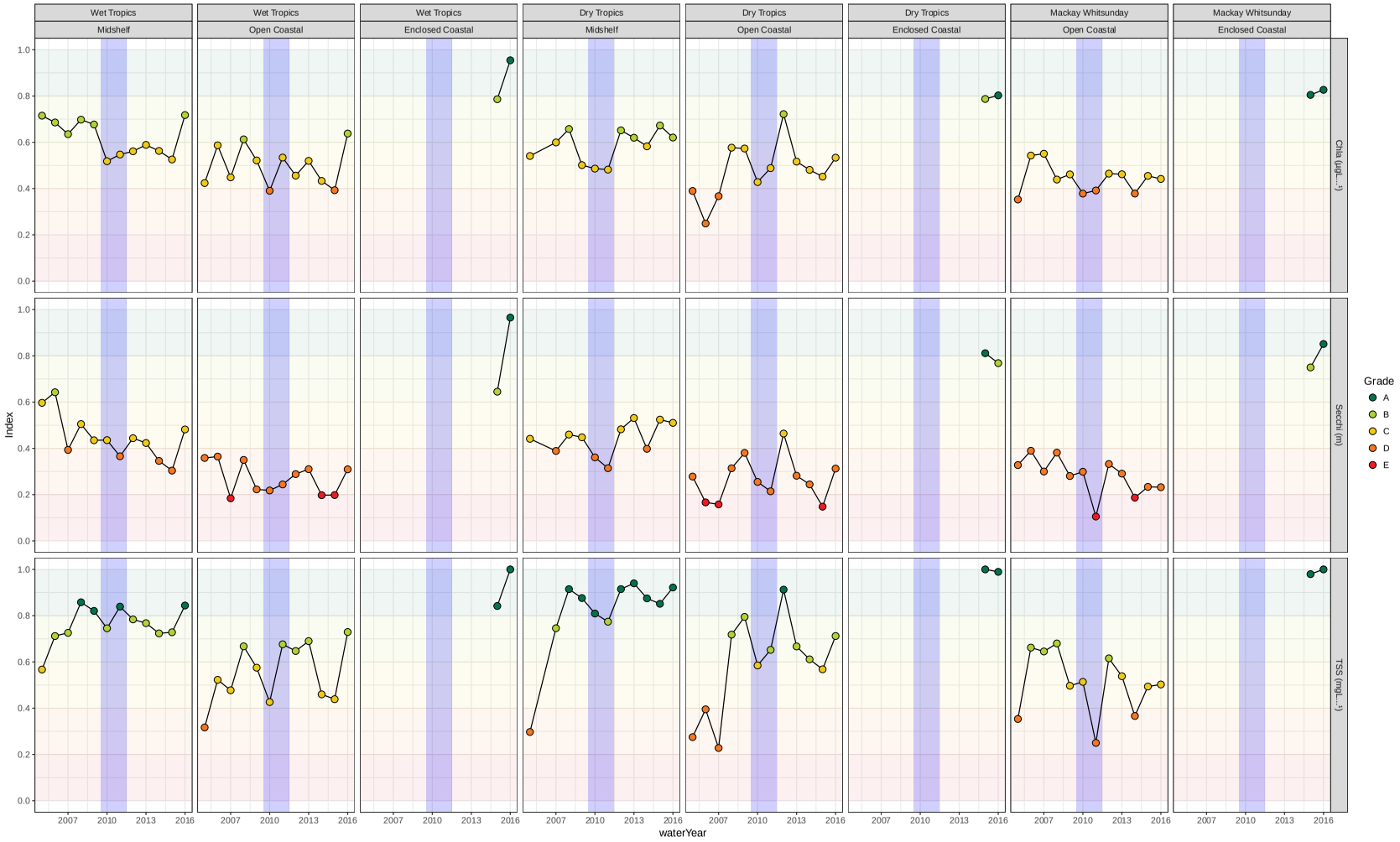


Figure 39: Time series of fsMAMP4 index (Seasonal thresholds) scores by zone. The blue vertical bar spans from mid 2009 to mid 2011.

Bootstrapping

Measure hierarchy last (Figure 7, Option d)

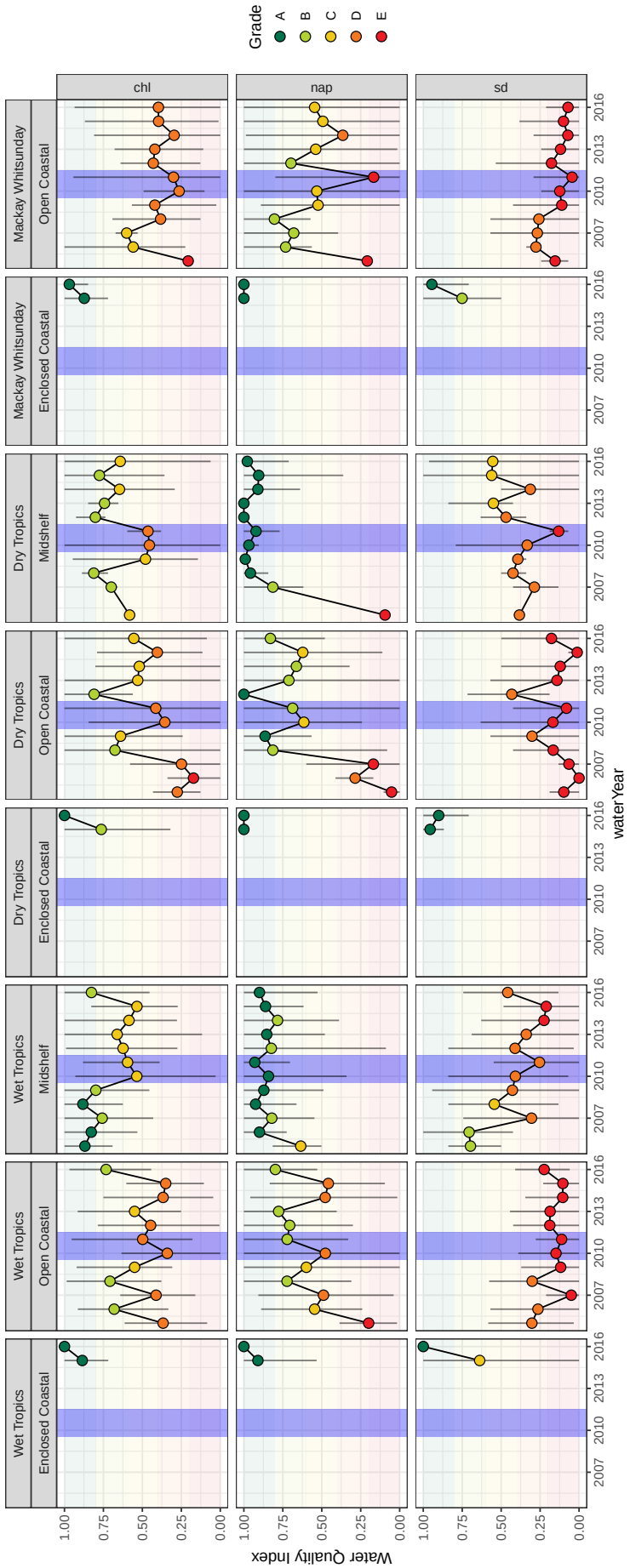


Figure 40: Time series of fsMAMP Measure index (Seasonal thresholds) scores by zone. The blue vertical bar spans from mid 2009 to mid 2011.

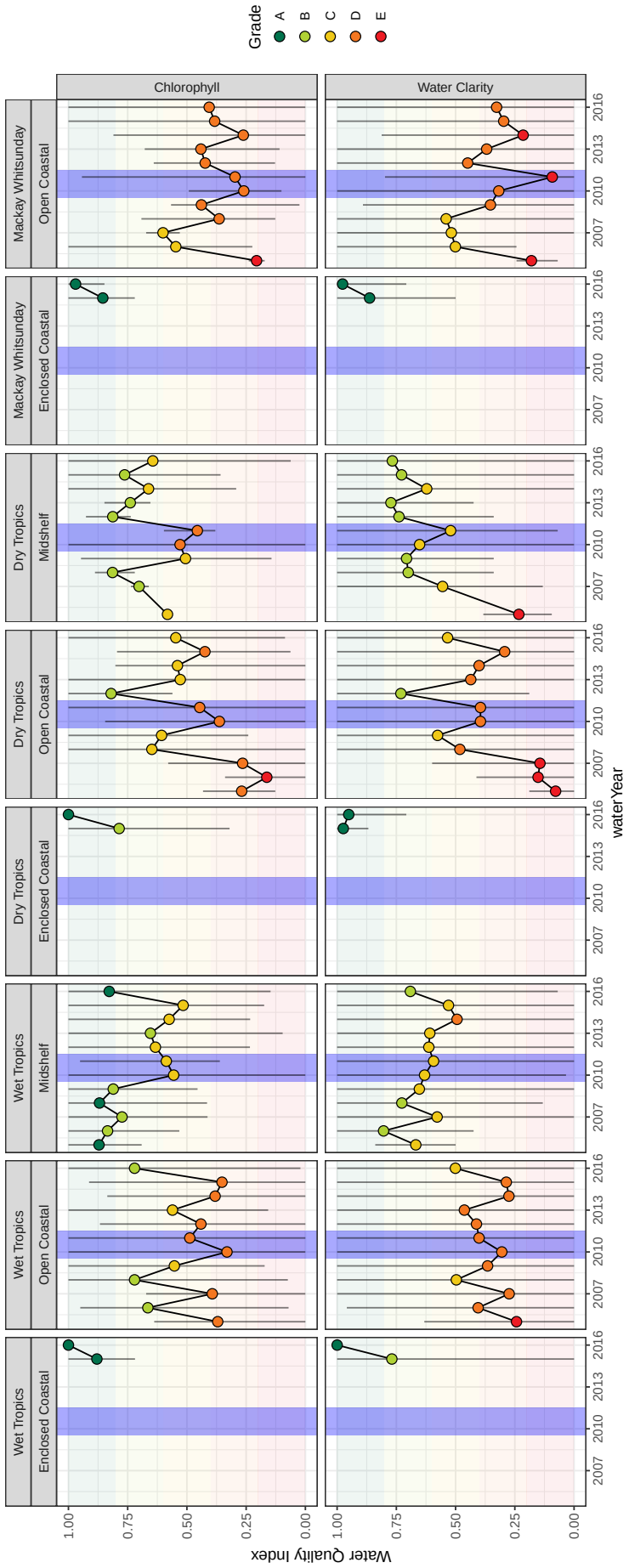


Figure 41: Time series of fsMAMP Subindicator index (Seasonal thresholds) scores by zone. The blue vertical bar spans from mid 2009 to mid 2011.

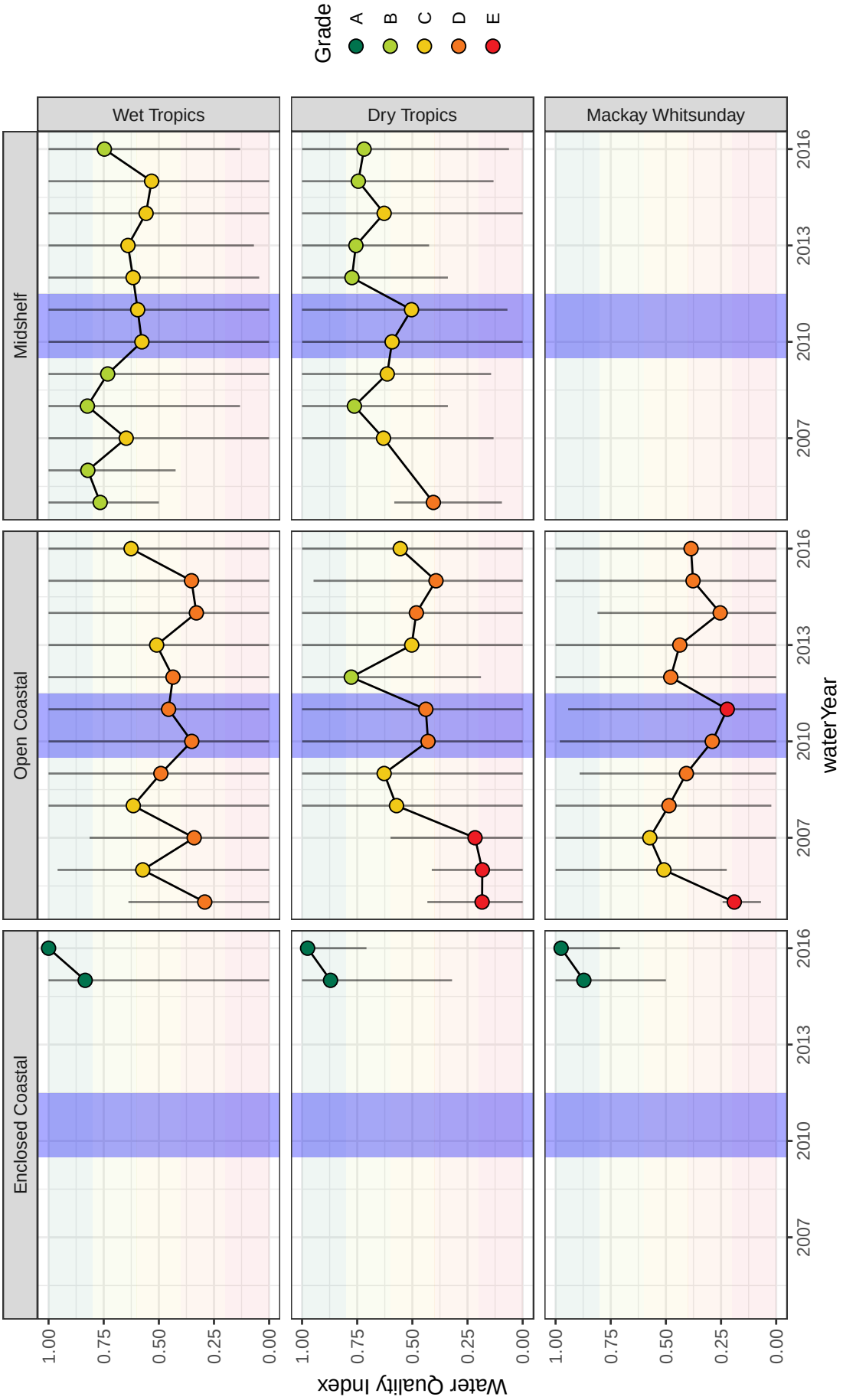


Figure 42: Time series of fsMAMP Indicator index (Seasonal thresholds) scores by zone. The blue vertical bar spans from mid 2009 to mid 2011.

Spatial hierarchy last (Figure 7, Option c)

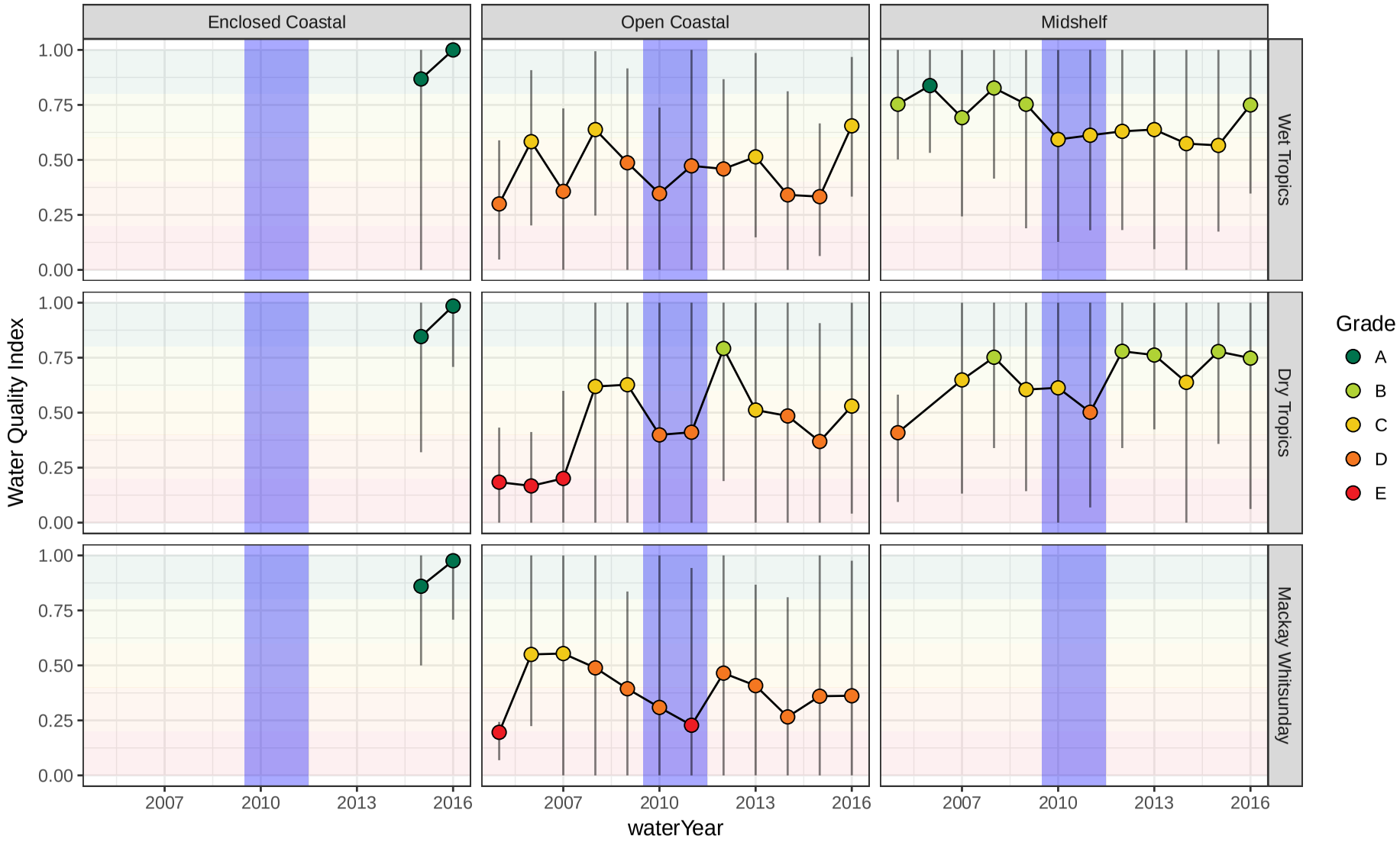


Figure 43: Time series of fsMAMP Indicator index (Seasonal thresholds) scores by zone. The blue vertical bar spans from mid 2009 to mid 2011.

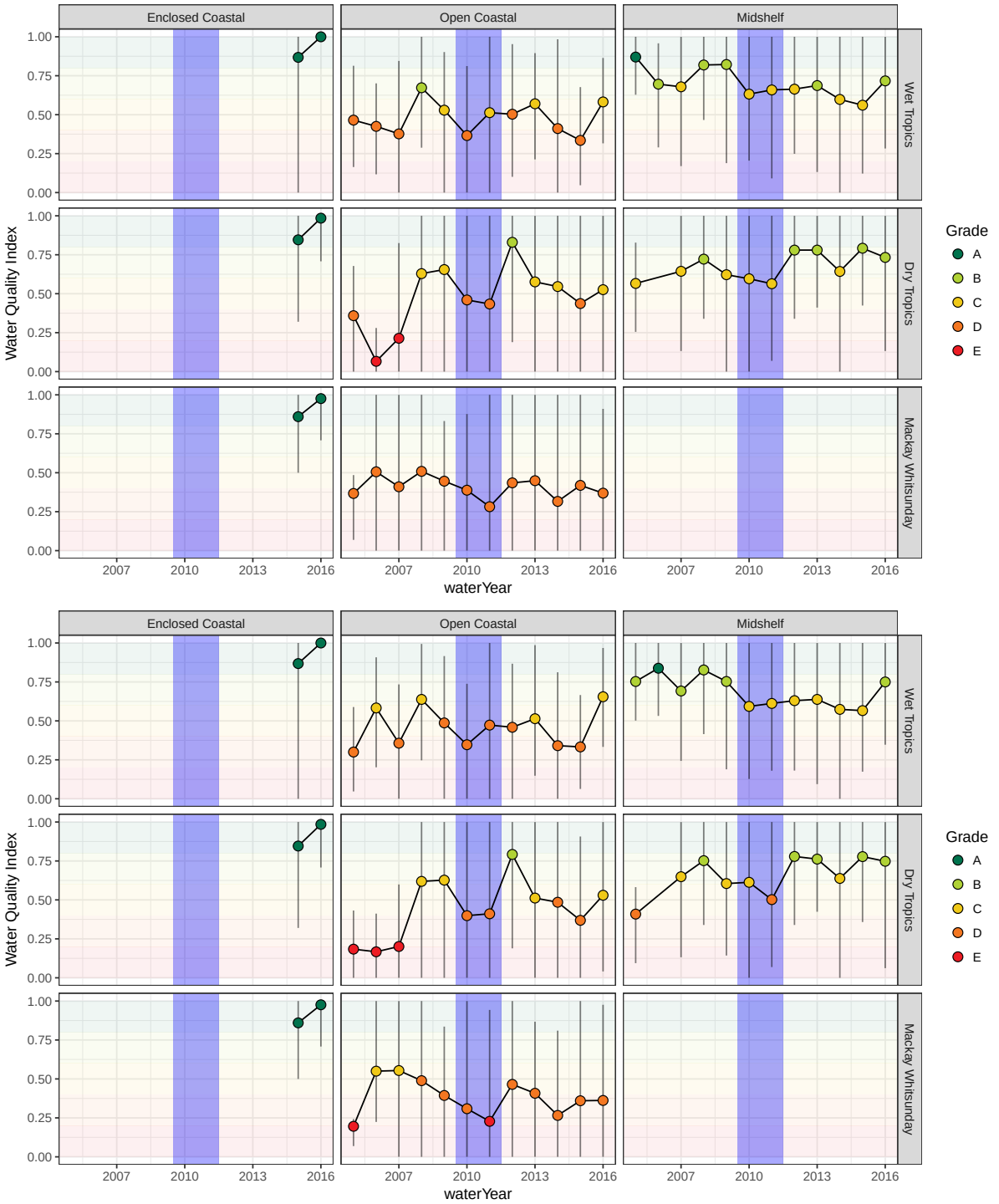


Figure 44: Comparison of time series of fsMAMP Indicator index (Annual vs Seasonal thresholds) scores by zone. The blue vertical bar spans from mid 2009 to mid 2011. The top figure is based on Annual thresholds and the bottom figure is based on Seasonal thresholds. This a comparison of outcomes from Figure 7 Options b and d.