Baytrends Cluster Analysis Methods.

Results from James River

Data Integrity Work Group
Sept. 21, 2022

presented by

Elgin Perry
Statistics Consultant
Riverside Farm, Leedstown, Va.

On Behalf of the

Integrated Trends Assessment Team

Road Map:

1. Quick Review of *baytrends* methods and results.

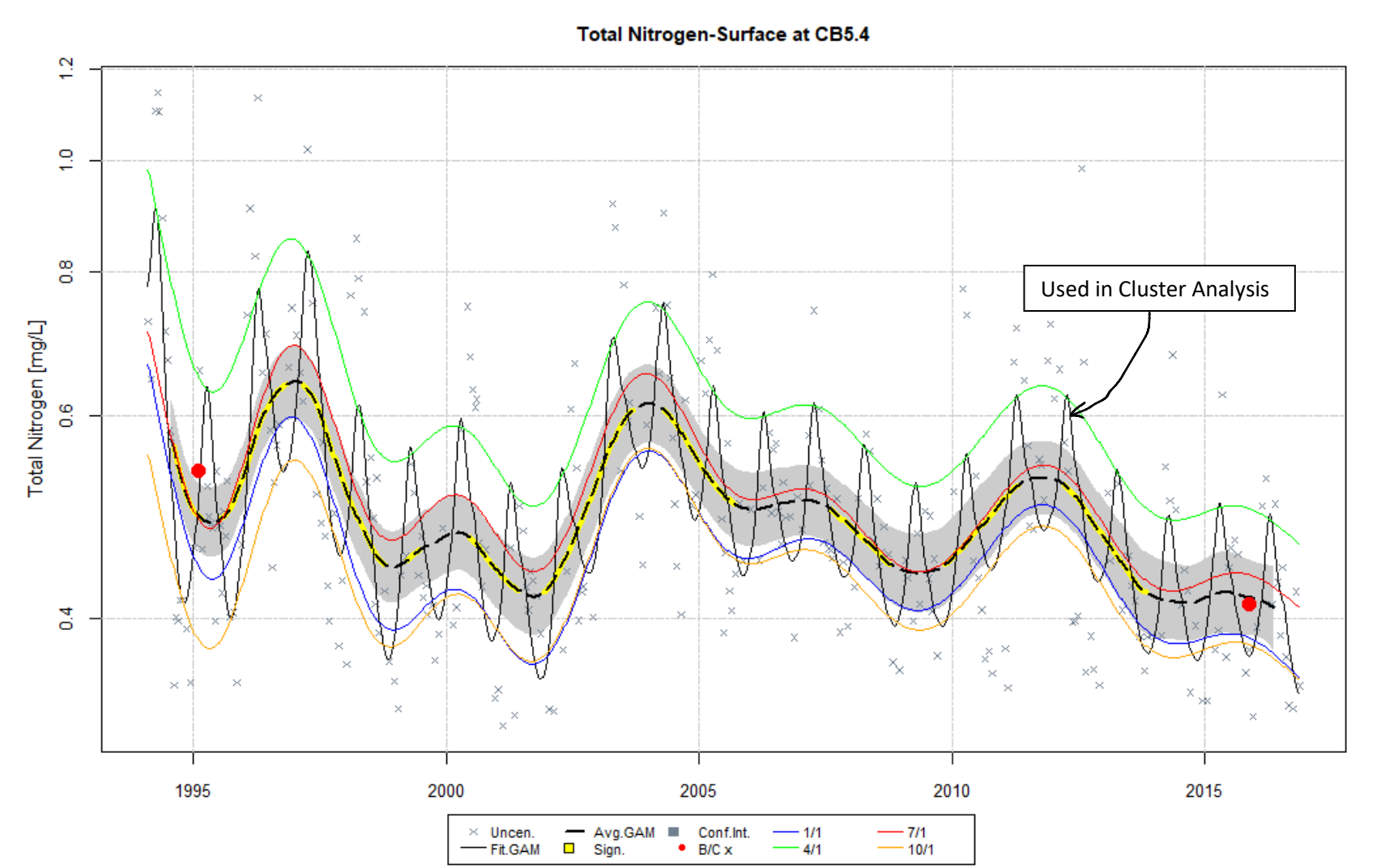
2. Motivation for applying Cluster Analysis to *baytrends* results

3. *Baytrends* Cluster Analysis Methods

4. Overview of types of cluster results

5 Example from Virginia Tidal Network.

6. Plans for the Future



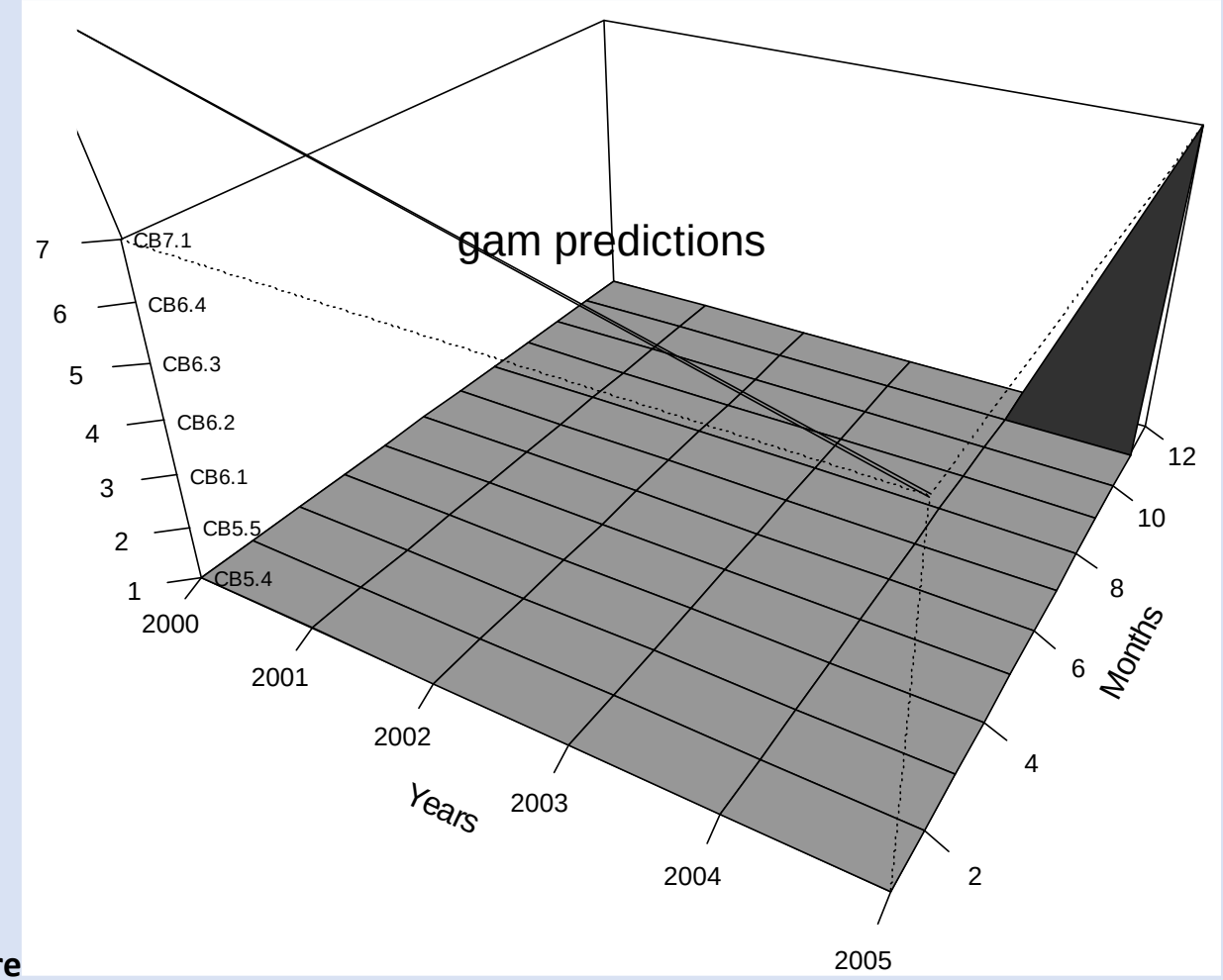
Generalized Additive Model (gam) Terms:

Log(TN) = Intercept + s(longTermTrend)+s(season) + interaction.smooth(Trend x Season)

Cluster Data:

Using the gam, obtain an estimated value for the

1. **15th of each month,**
2. **for each year in the period of record**
3. **for each station in the area of interest.**



Store these as a 3-D data Structure

Convert to 2-D Data Structure of Items and Profiles.

		Profile									
	station	tn.pred 1994	tn.pred 1995	tn.pred 1996	tn.pred 1997	tn.pred 1998	tn.pred 1999	tn.pred 2000	tn.pred 2001	tn.pred 2002	tn.pred 2003
ITEMS	CB5.4	-0.5924	-0.7001	-0.5208	-0.5340	-0.7558	-0.7856	-0.7662	-0.8663	-0.8072	-0.5796
	CB5.4W	-0.5698	-0.5781	-0.4277	-0.4764	-0.6667	-0.7067	-0.7119	-0.7790	-0.7073	-0.5162
	CB5.5	-0.5845	-0.7292	-0.5950	-0.5848	-0.7794	-0.8397	-0.8399	-0.9113	-0.8342	-0.6269
	CB6.1	-0.6238	-0.7275	-0.6123	-0.6204	-0.7982	-0.8416	-0.8350	-0.9157	-0.8451	-0.6210
	CB6.2	-0.6888	-0.7855	-0.6552	-0.6591	-0.8368	-0.8760	-0.8629	-0.9264	-0.8433	-0.6405
	CB6.3	-0.7045	-0.7957	-0.6475	-0.6658	-0.8572	-0.8808	-0.8645	-0.9545	-0.8921	-0.6796
	CB6.4	-0.7317	-0.8371	-0.7771	-0.7797	-0.9529	-1.0090	-0.9771	-1.0359	-0.9908	-0.7830
	CB7.1	-0.6817	-0.7944	-0.6825	-0.6673	-0.8290	-0.8834	-0.8751	-0.9104	-0.8241	-0.6555
	CB7.1N	-0.6351	-0.7352	-0.5558	-0.5933	-0.8223	-0.8319	-0.8008	-0.9063	-0.8424	-0.6159
	CB7.1S	-0.7368	-0.8463	-0.6737	-0.6671	-0.8741	-0.9217	-0.8917	-0.9318	-0.8454	-0.6604

Baytrends Cluster Methods

- **Agglomerative Clustering**
- **Ward's Method to Form Clusters**
- **Dendogram to display tree of grouping structure**
- **Profile plots to assess characteristics of each group**
- **User chooses Years, Months, and Stations to analyze.**
- **User defines Items and Profiles by choosing from: Years, Months, and Station**
- **User chooses scaling of Profile**
- **User chooses number of Groups to interpret**
- **New Features: Automatic Group Labels, Auto Color Assignments**

Variations on Clustering

Item	Profile	Scaling	Question
Station	Year	None	What stations have similar long term means?
Station	Year	Mean Adjust	What stations have similar long term trends? How does Water Quality respond to Flow?
Station	Month	Mean Adjust	What stations have similar seasonal patterns?
Station:Year	Month	None	Does Seasonality change over Years within Station?
Year	Station	None	Does the Estuary profile differ between high flow and low flow years? Is trend consistent across the estuary?
Month	Station	None	On average over years, how does upstream-downstream profile change with season?

1. **Status Cluster: Items:Station Profile:Year Scale: None**

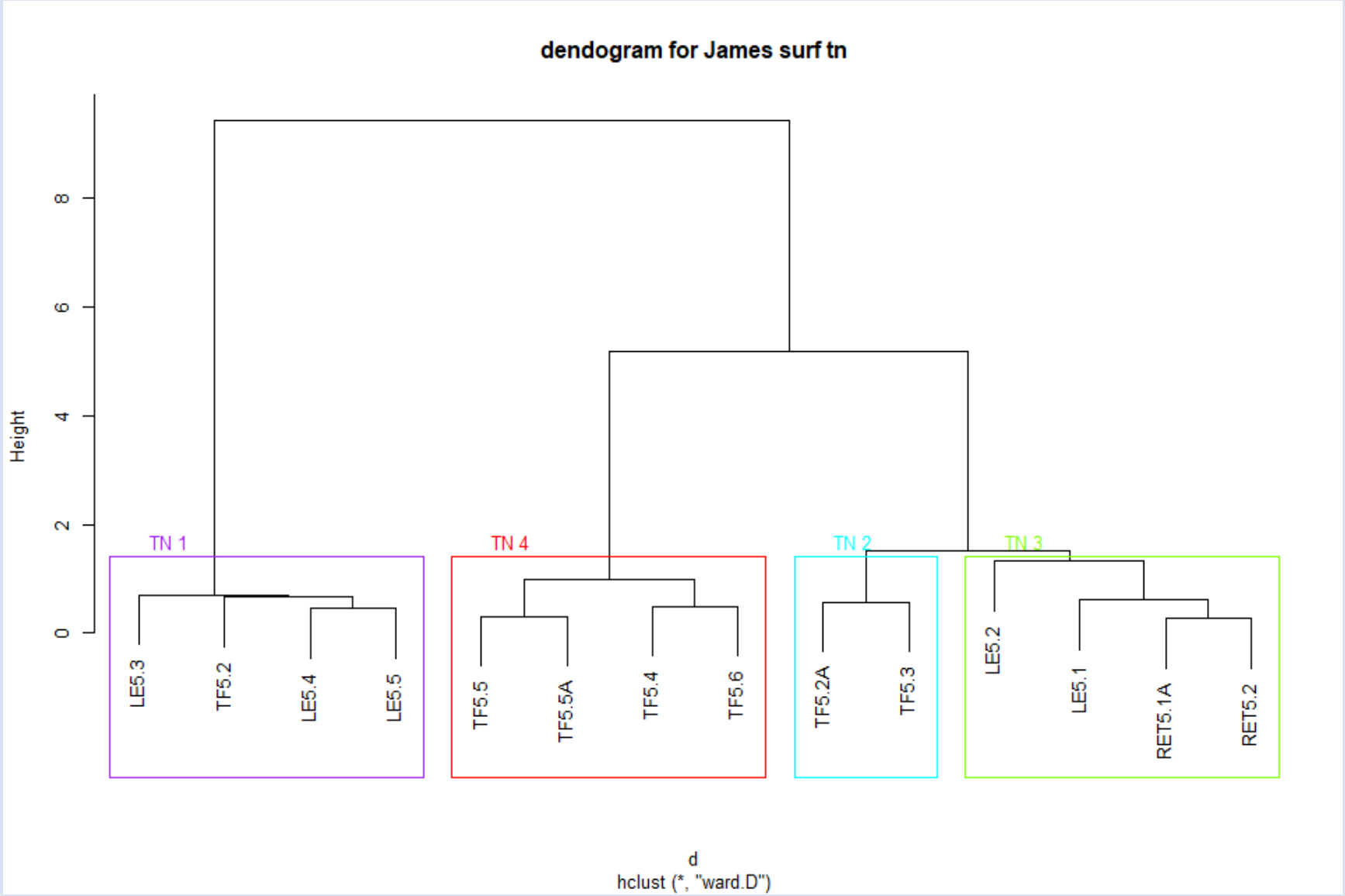


Figure 3.A: Dendogram showing distances between station pairs computed using the dist() function for James River.

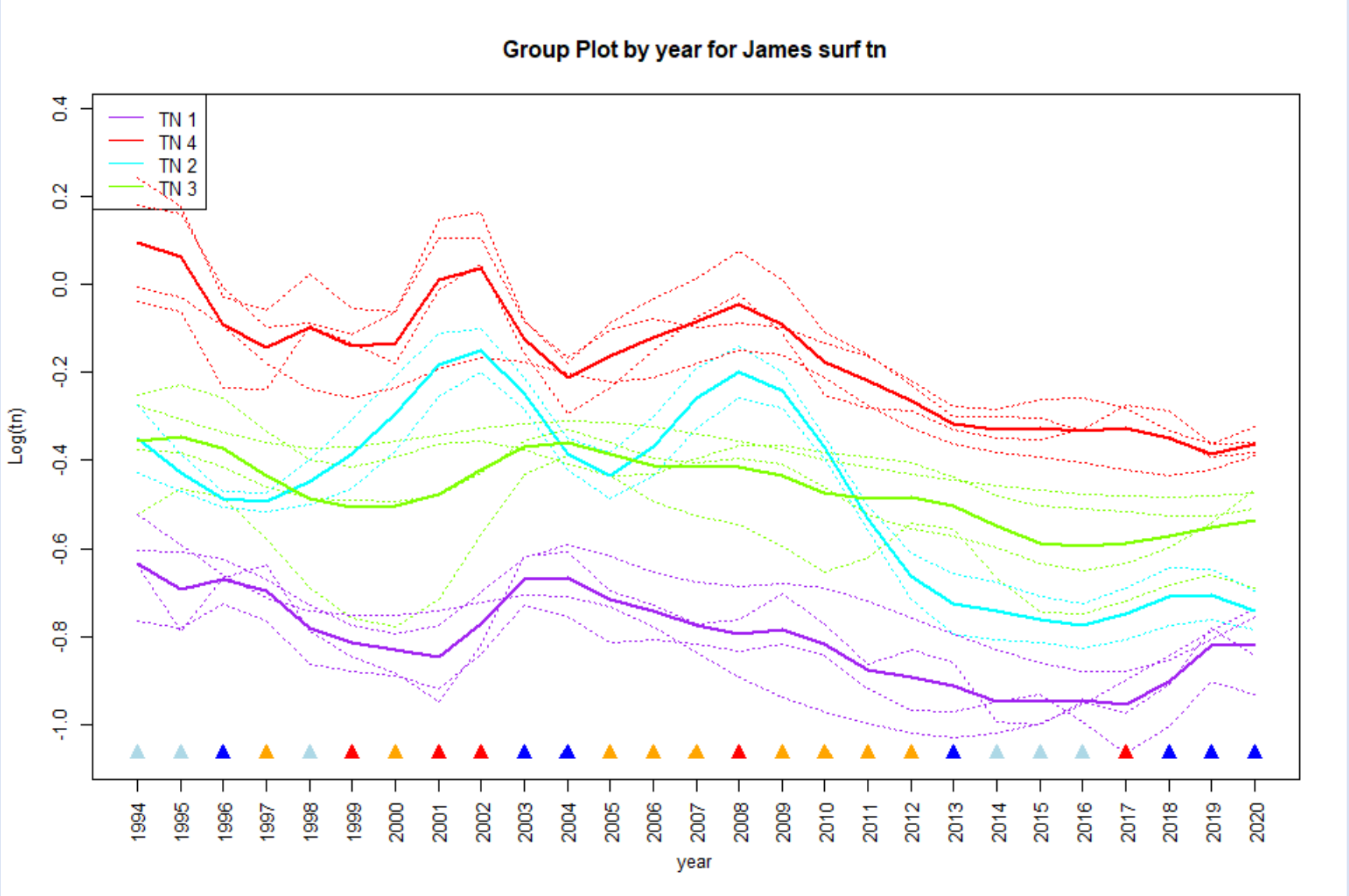
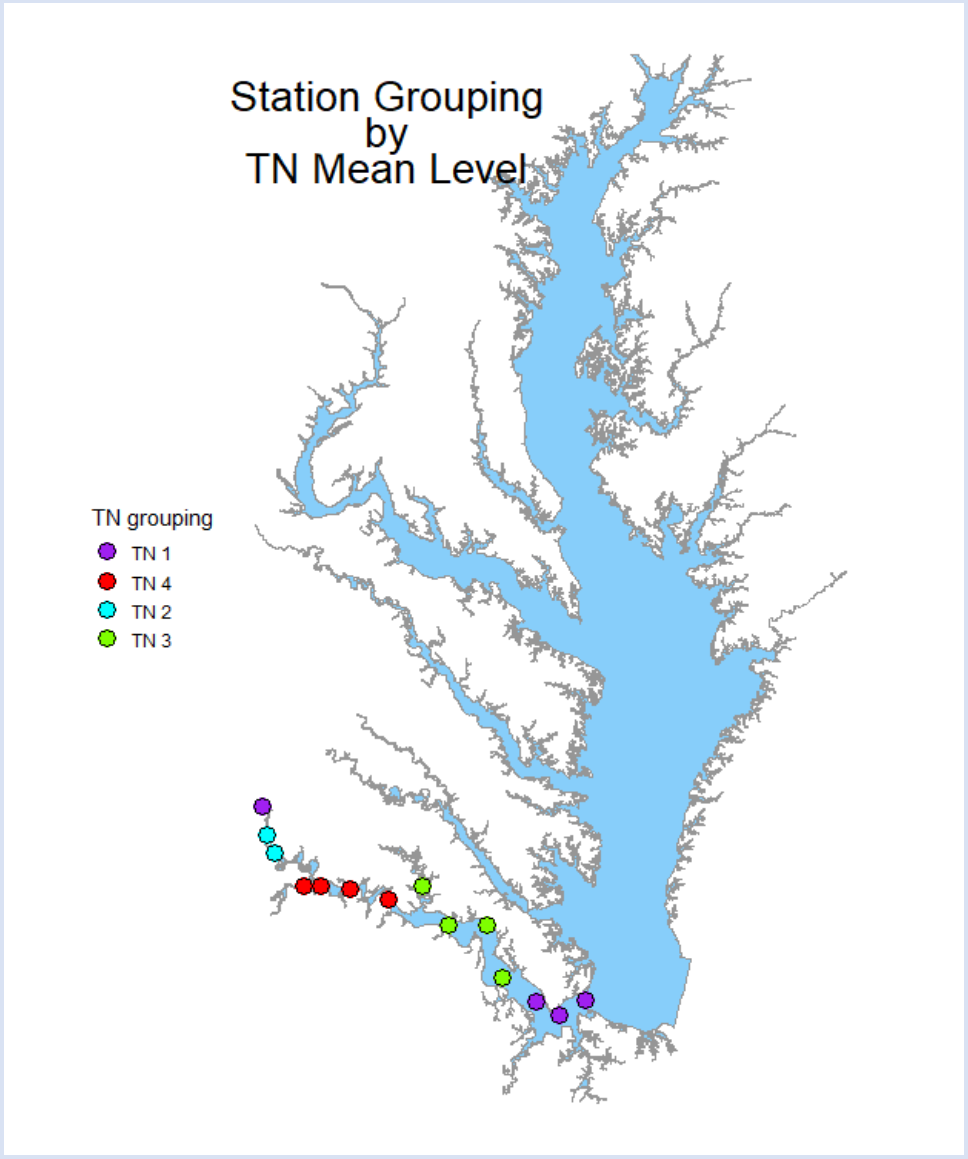
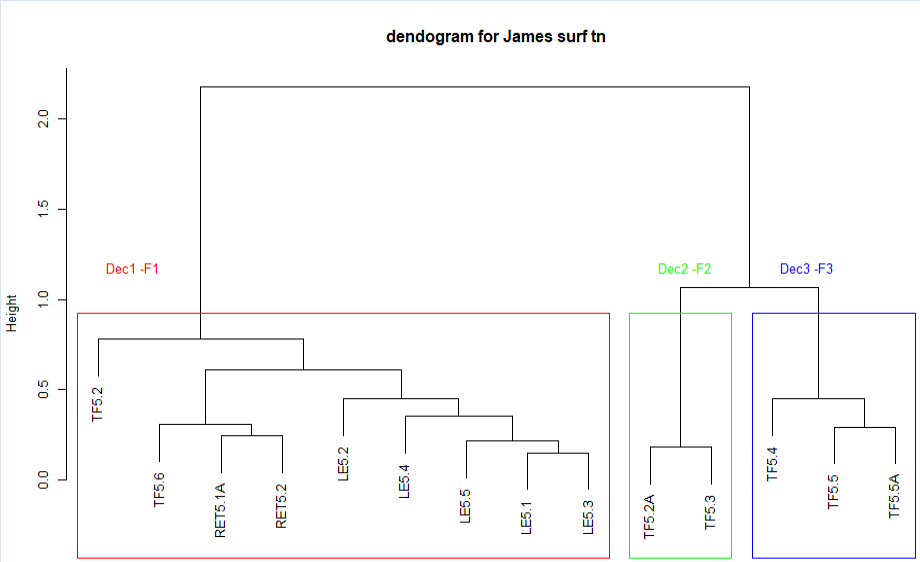
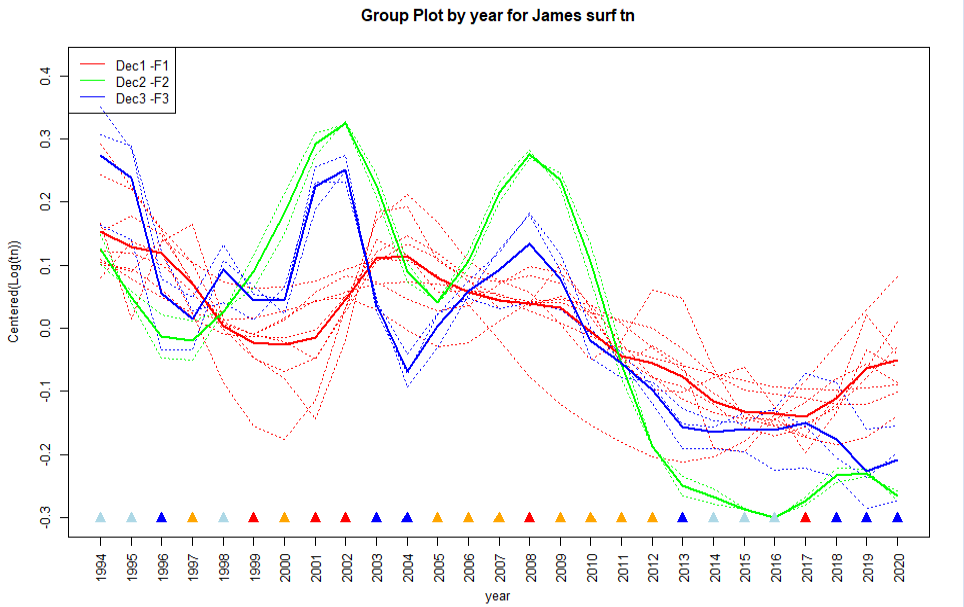


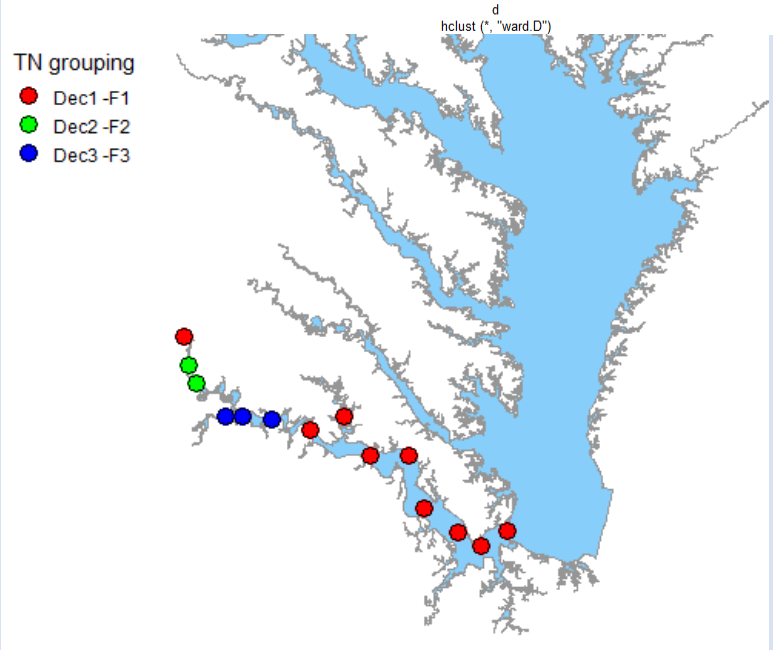
Figure 3.B: Year means plotted with station groups segregated by color. Multiple dash line traces within group show variability among station within groups. Items have been mean adjusted.


Station Grouping
by
TN Mean Level
TN grouping
TN 1
TN 4
TN 2
TN 3

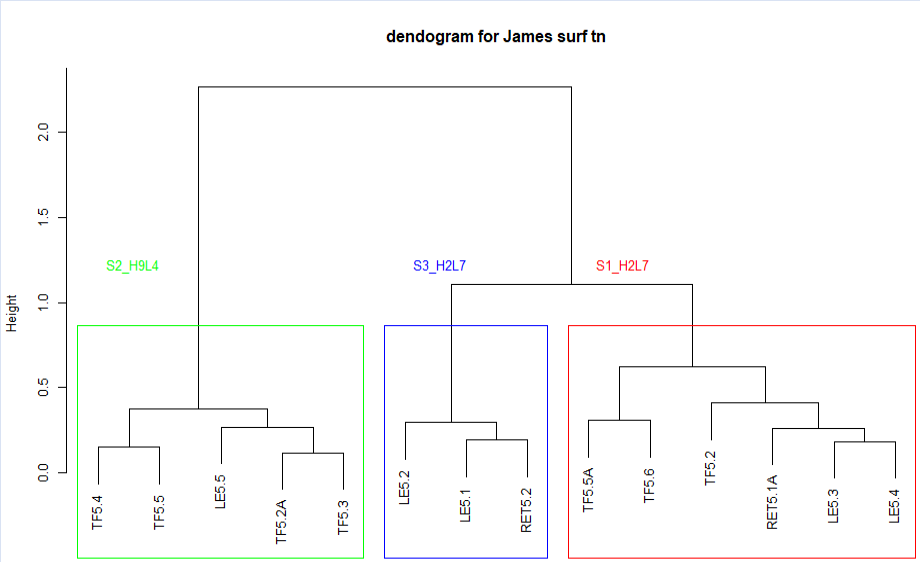
2. Trends Cluster: Items:Station Profile: Year Scale: Mean Adjusted

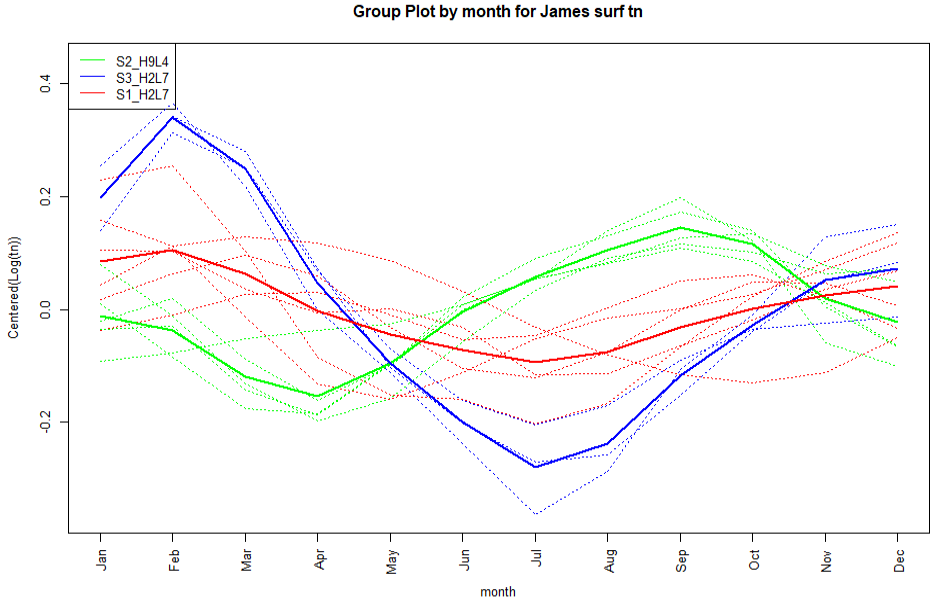


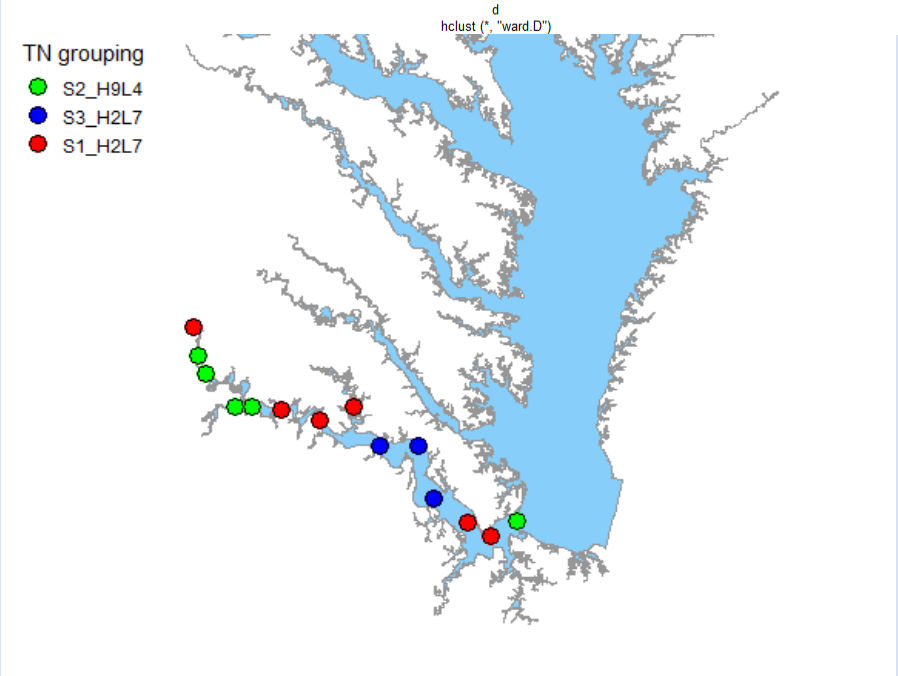




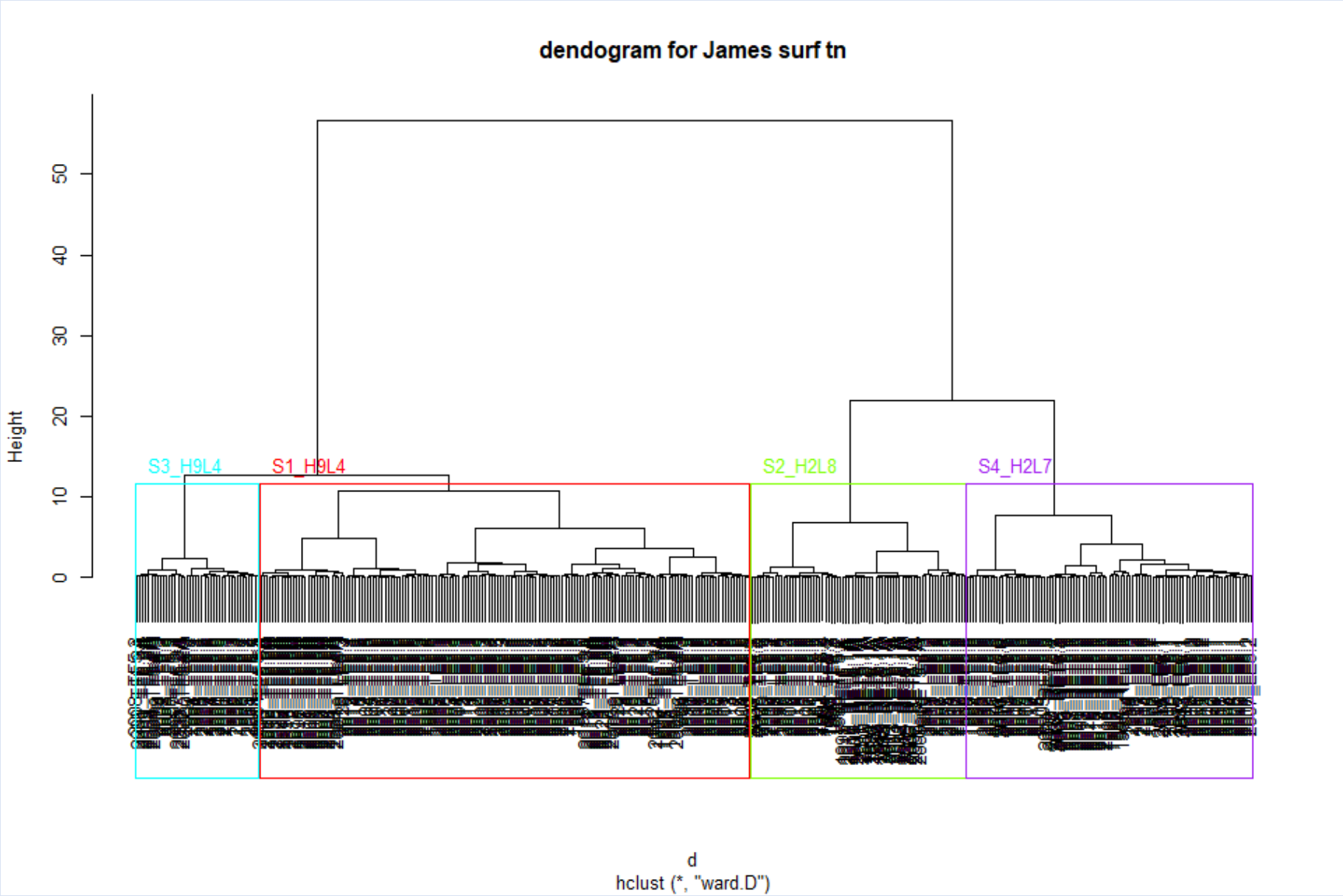
3. Seasons Cluster: Items:Station Profile: Month Scale: Mean Adjusted

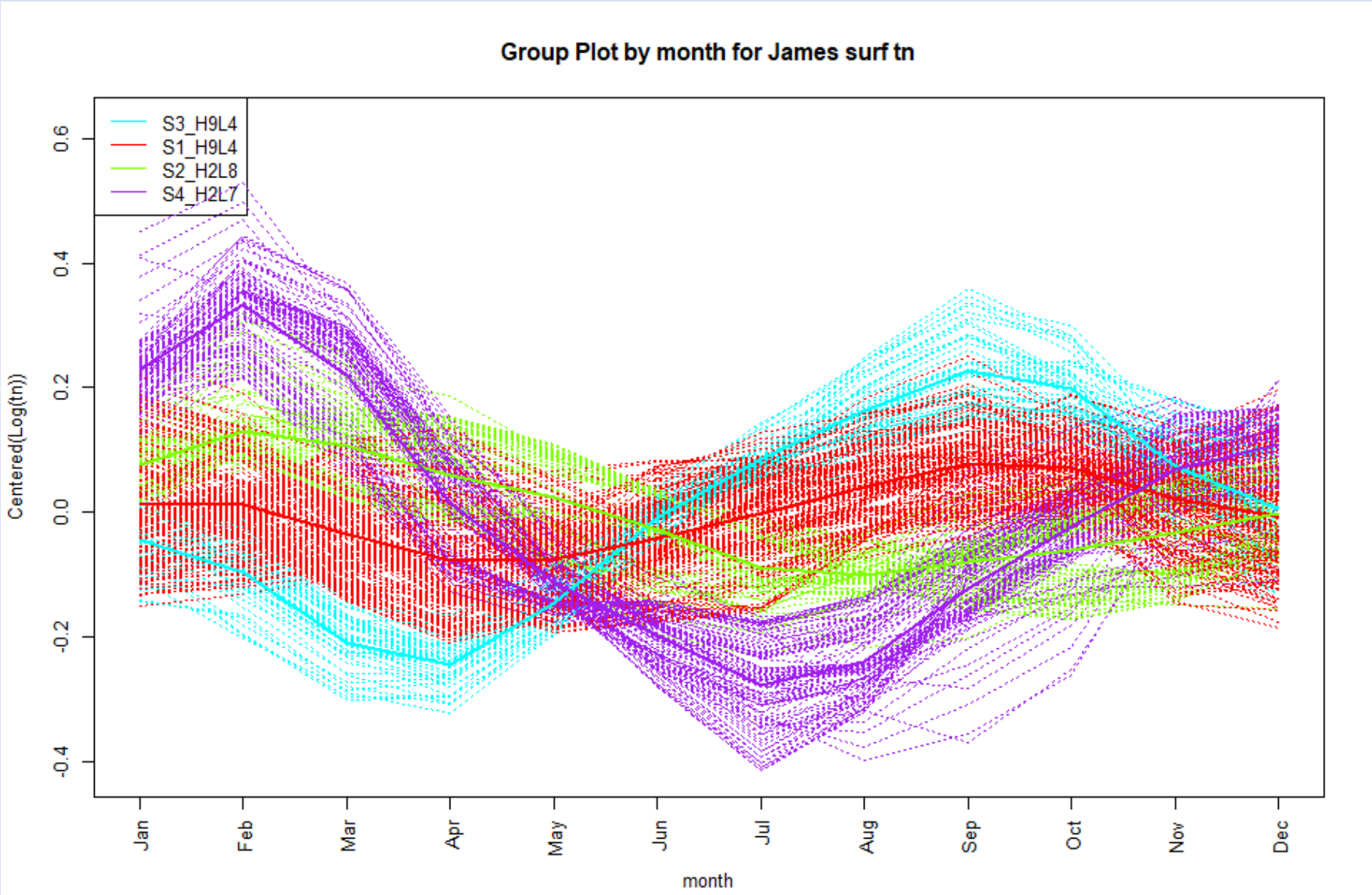


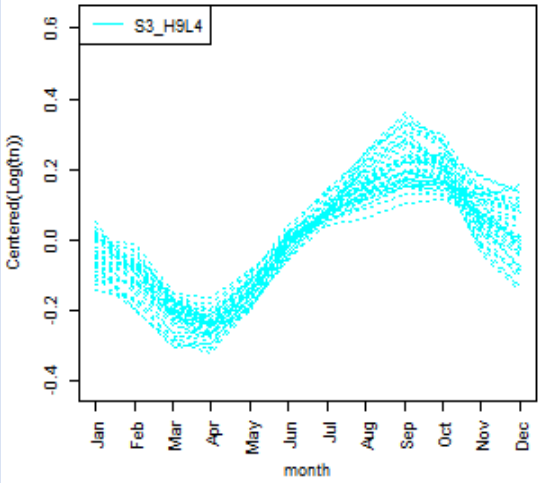


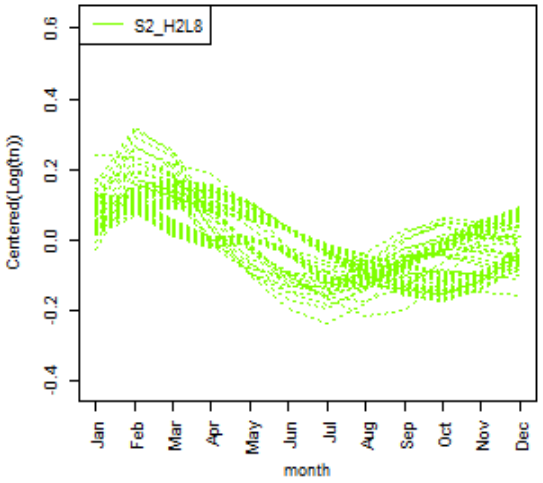


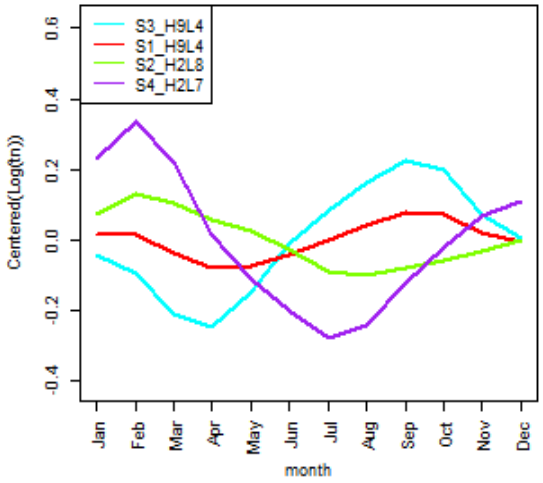
4. Climate Change Cluster: Items:Station-Year Profile: Month Scale: Mean Adjusted

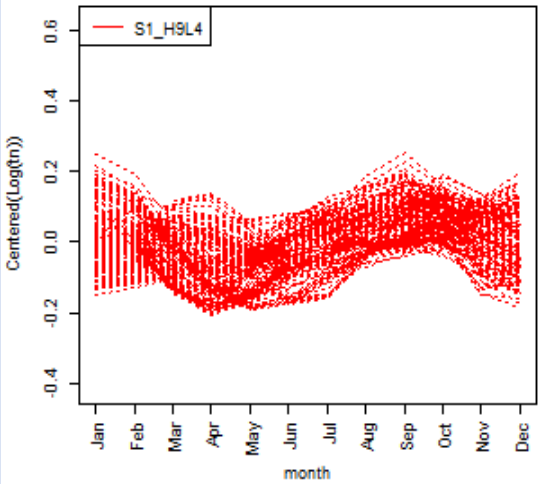


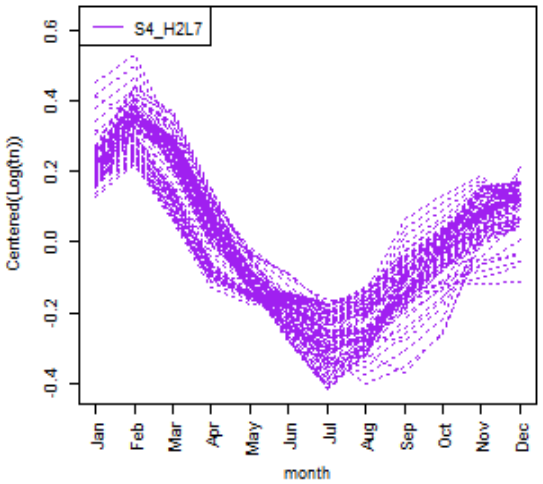

Group Plot by month for James surf tn
S3_H9L4
S1_H9L4
S2_H2L8
S4_H2L7
Centered(Log(tn))
0.6
0.4
0.2
0.0
-0.2
-0.4
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
month

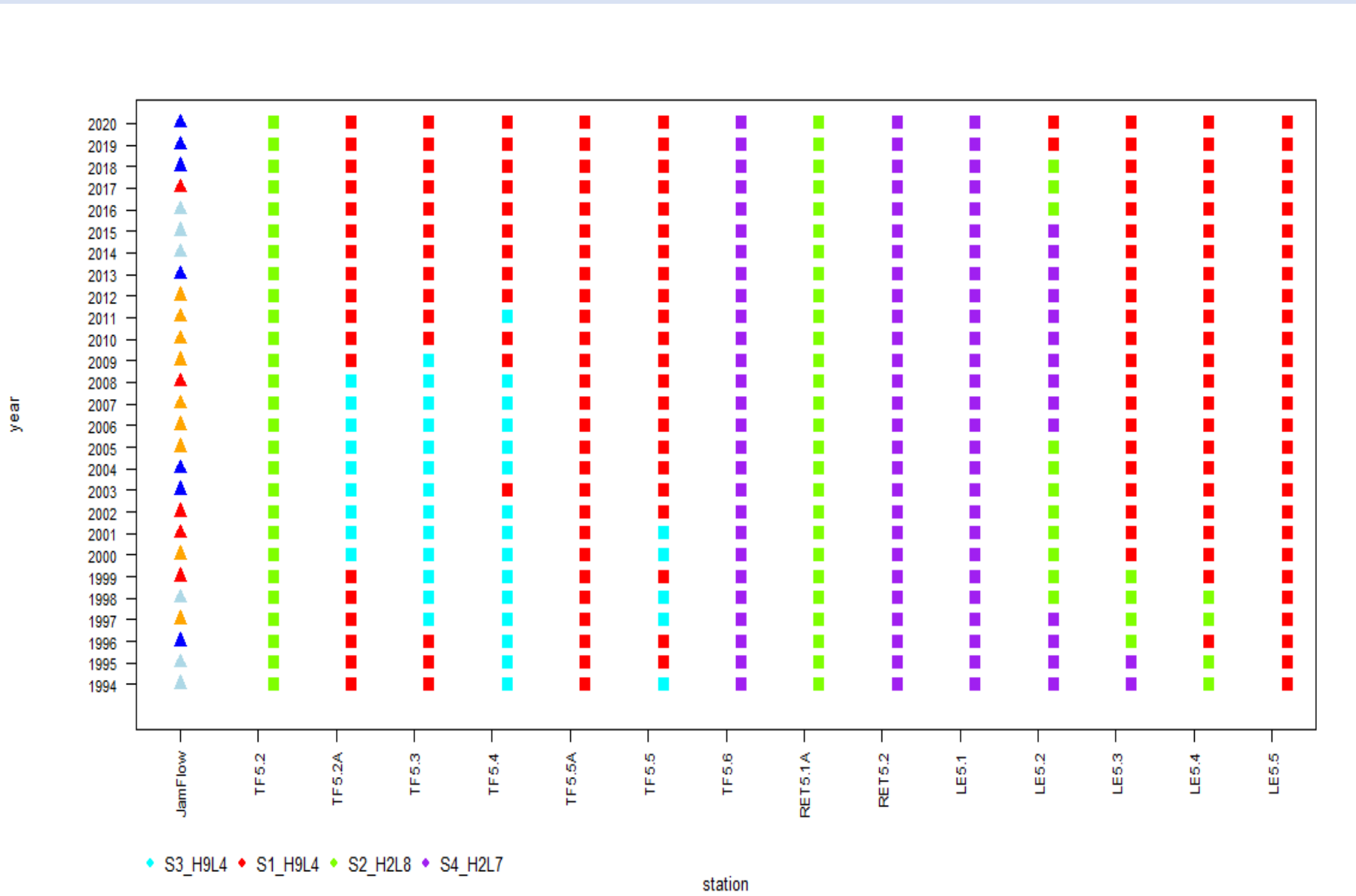

S3_H9L4
Centered(Log(tn))
-0.4
-0.2
0.0
0.2
0.4
0.6
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
month


S2_H2L8
Centered(Log(tn))
-0.4
-0.2
0.0
0.2
0.4
0.6
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
month

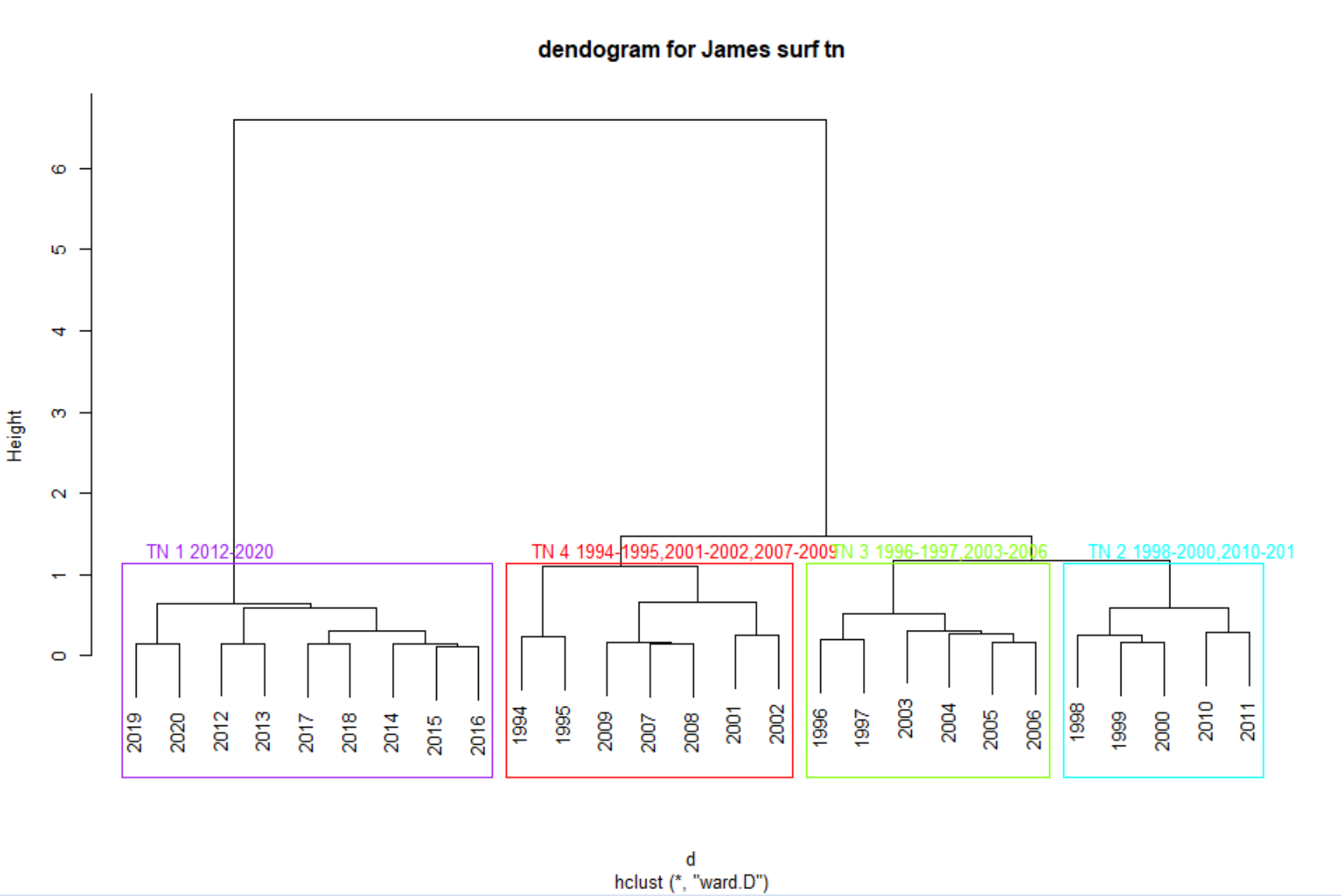

S3_H9L4
S1_H9L4
S2_H2L8
S4_H2L7
Centered(Log(tn))
-0.4
-0.2
0.0
0.2
0.4
0.6
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
month

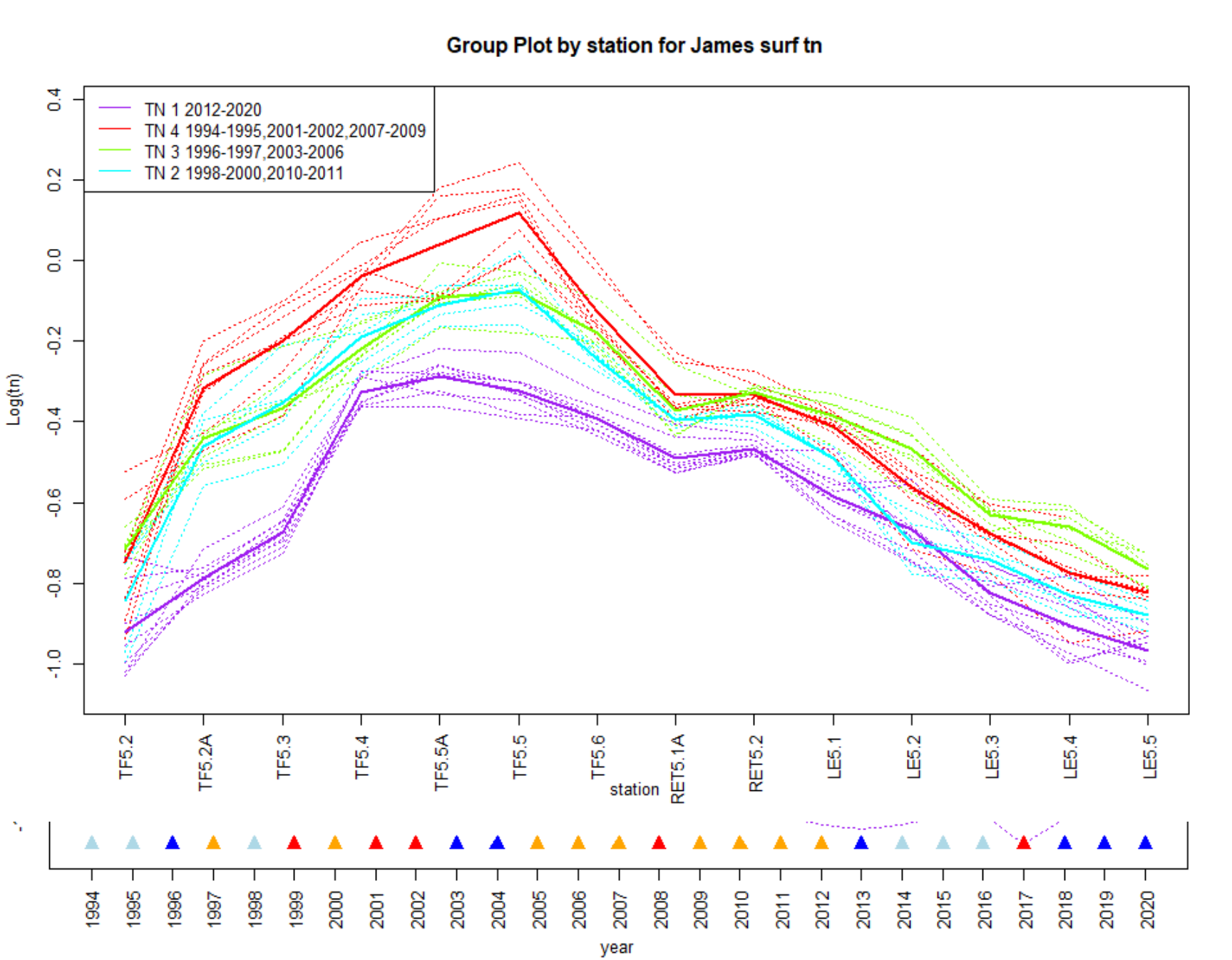

S1_H9L4
Centered(Log(tn))
-0.4
-0.2
0.0
0.2
0.4
0.6
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
month


S4_H2L7
Centered(Log(tn))
-0.4
-0.2
0.0
0.2
0.4
0.6
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
month


year
2020
2019
2018
2017
2016
2015
2014
2013
2012
2011
2010
2009
2008
2007
2006
2005
2004
2003
2002
2001
2000
1999
1998
1997
1996
1995
1994
JamFlow
TF5.2
TF5.2A
TF5.3
TF5.4
TF5.5A
TF5.5
TF5.6
RET5.1A
RET5.2
LE5.1
LE5.2
LE5.3
LE5.4
LE5.5
S3_H9L4
S1_H9L4
S2_H2L8
S4_H2L7
station

5. **Estuarine Profile Cluster: Items:Year Profile: Stations Scale: none**




Group Plot by station for James surf tn
TN 1 2012-2020
TN 4 1994-1995,2001-2002,2007-2009
TN 3 1996-1997,2003-2006
TN 2 1998-2000,2010-2011
Log(tn)
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8
-1.0
TF5.2
TF5.2A
TF5.3
TF5.4
TF5.5A
TF5.5
TF5.6
RET5.1A
RET5.2
LE5.1
LE5.2
LE5.3
LE5.4
LE5.5
station
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
year

Advantages of baytrends Cluster Approach:

- **Organizes GAMs results to tell a Status and Trends story.**
- **Allows data to self-organize.**
- **Test existing concepts for explaining trends.**
- **Explore new Concepts (e.g. changing seasonality, spatial trends)**

Concerns:

- **Gam smoothing may de-emphasize differences between neighboring years**
- **Auto-labeling may interfere with critical thinking.**
- **??**

Plans for the Future:

- **Incorporate Cluster Results in Tributary Summary Reports**
- **Develop Cluster algorithms as add-on package to baytrends**
- **Clustering of flow GAMs, with and without flow adjustment**
- **Add Mixing Curves (Water Quality vs. Salinity)**
- **Allow for 2 term profiles**
- **Add more scaling options (percent, z-score, exp)**
- **Add facility to compare groups among clusters.**
- **Use CART to explain groupings based on watershed features.**
- **Add Water Quality parameter are 4^{th} dimension of data structure.**

Sunset on the Rappahannock at Riverside Farm.