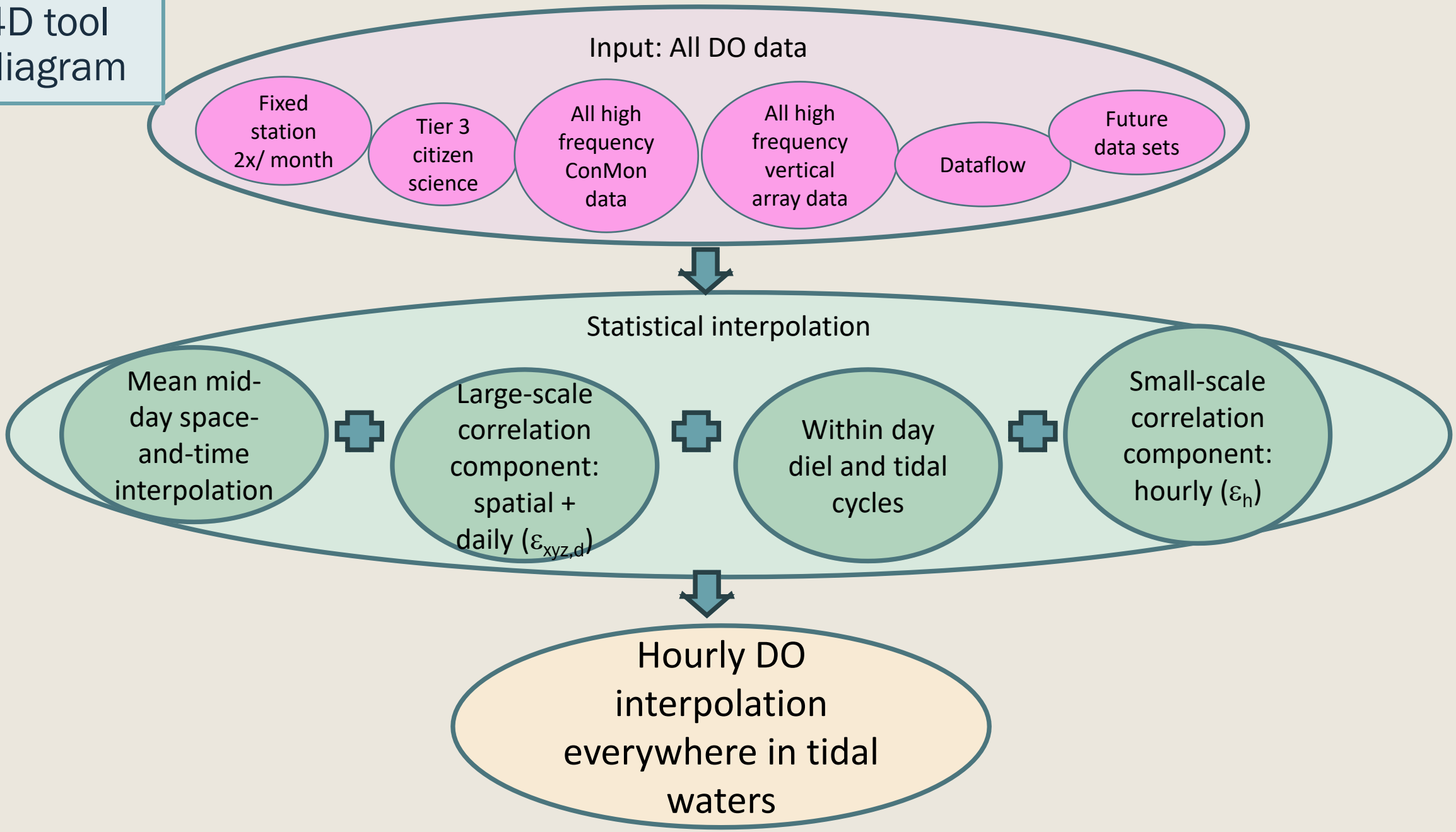
Class Imbalance and Proposed Experimental Units

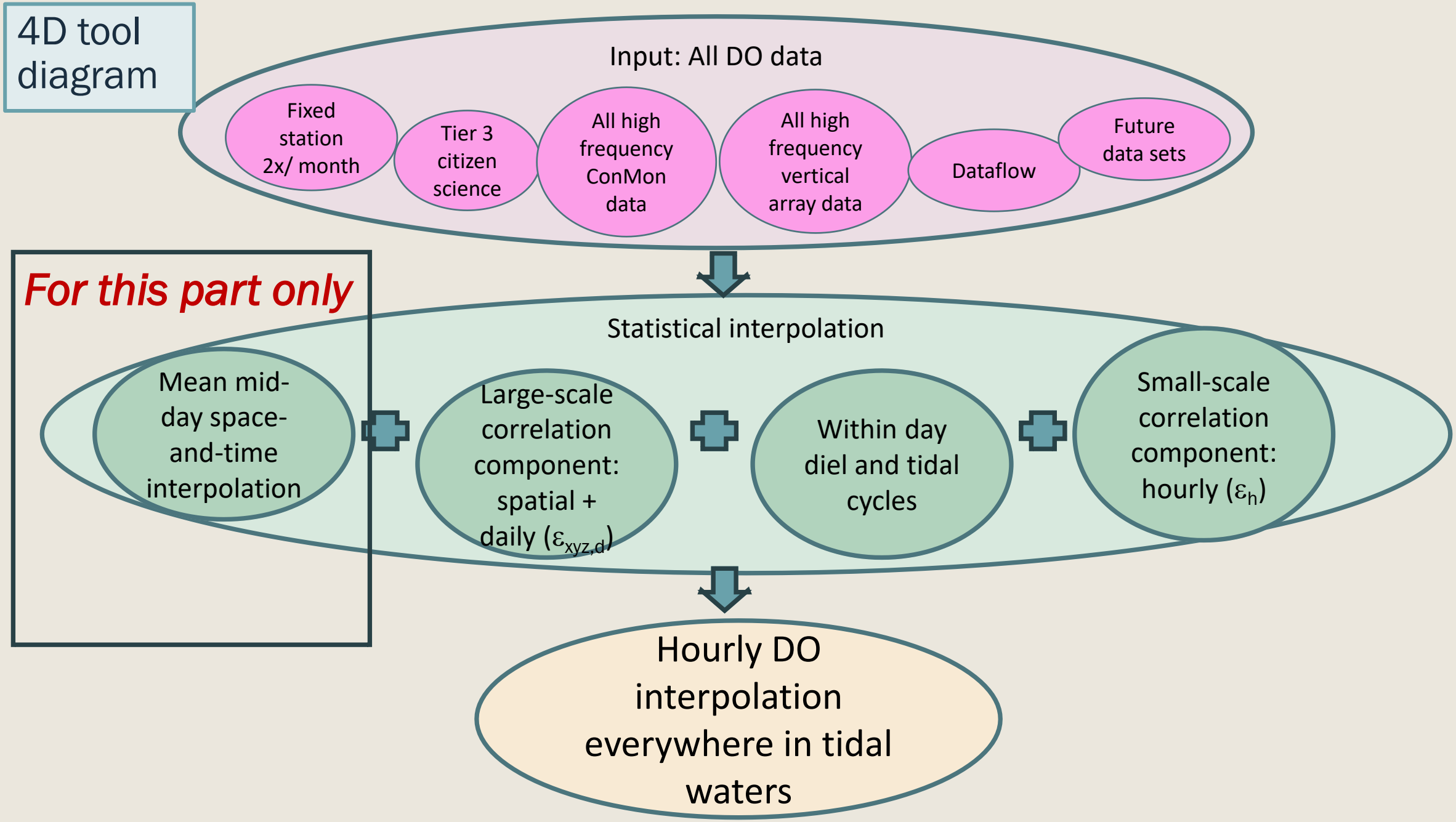
BORG meeting
Jan. 26, 2026

Rebecca Murphy (UMCES at CBP) and Elgin Perry (Statistics Consultant)

4D tool diagram



4D tool diagram



Introduction

- A 4-D interpolator strength and challenge is using data sets with very different spatial and temporal scales together.
  - *High frequency temporal data tends to be in shallow waters and can far exceed the count of boat-collected (grab samples) in a year: >99% to <1%.*
  - ***The high frequency data is critical to many parts of the interpolation**. However, without careful consideration, shallow water DO patterns could be over-represented in the results.*
- The GAM approach was designed to deal with this, to some degree, by fitting relationships between location and DO. But it appears there is a limit to overcoming this data imbalance when data is input at a finer time step than daily.

Introduction

- A 4-D interpolator strength and challenge is using data sets with very different spatial and temporal scales together.
  - *High frequency temporal data tends to be in shallow waters and can far exceed the count of boat-collected (grab samples) in a year: >99% to <1%.*
  - ***The high frequency data is critical to many parts of the interpolation**. However, without careful consideration, shallow water DO patterns could be over-represented in the results.*
- The GAM approach was designed to deal with this, to some degree, by fitting relationships between location and DO. But it appears there is a limit to overcoming this data imbalance when data is input at a finer time step than daily.

Explorations have included:

Hourly timestep

In Nov., we presented the idea of using high frequency data at the hourly timestep. Analysis showed this does not lose any information on the shape of the DO distributions.

Insights from the literature

This is a well-known problem in many fields of data analysis called "**class imbalance**." We will not be outliers if we implement approaches to handle it. Elgin will discuss.

Experimental Units/Weighting

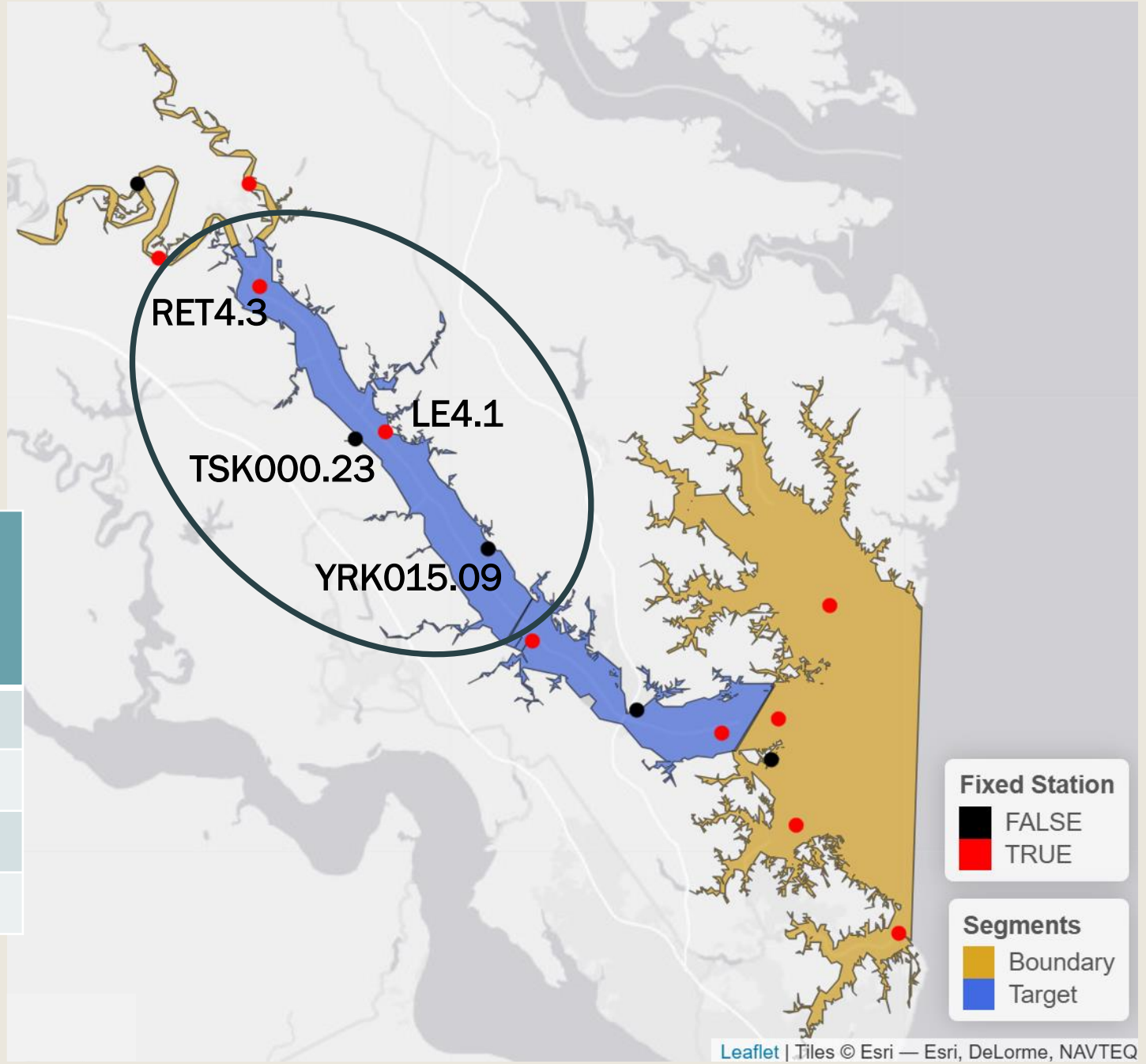
All hourly data will still be used in the GAM. However, different weights are applied to the frequency classes before fitting to balance the impact of high frequency-to-grab sample data.

Elgin’s slides

Example

2022 Stations in YRKMH

Station	Median bottom depth (m)*	Count of samples (hour-depths)*	Count of hours with samples*
RET4.3	7	56 (0.33%)	10 (0.06%)
LE4.1	10	87 (0.52%)	10 (0.06%)
TSK000.23	1.8	8,512 (51%)	8,512 (51%)
YRK015.09	1.9	8,146 (48%)	8,146 (49%)

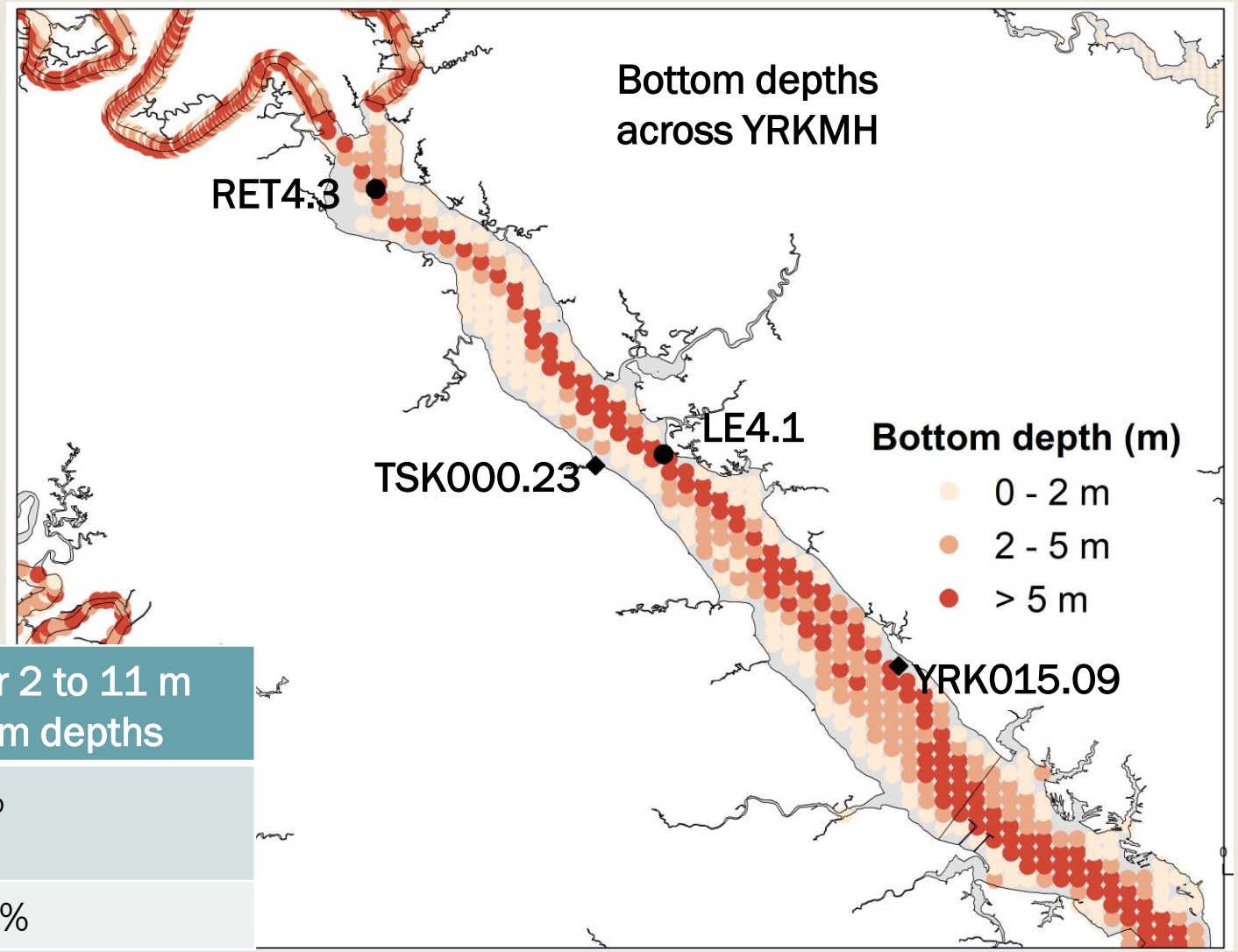


Spatial frequencies in sampling

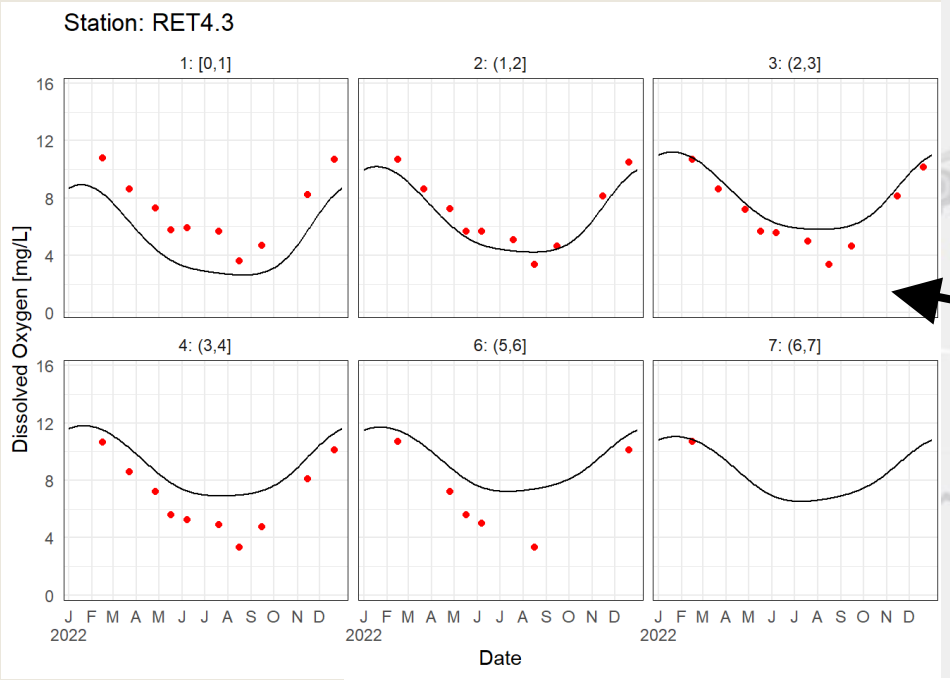
→ Shallow waters are over-represented in this data.

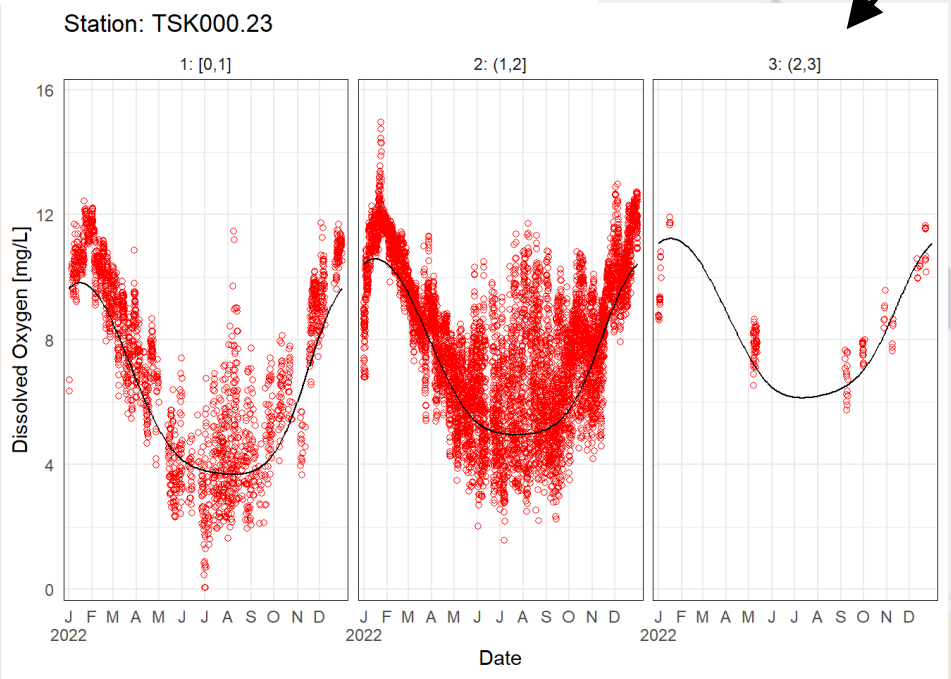
	Water 0 to 2m bottom depths	Water 2 to 11 m bottom depths
Percent of hourly samples*	99.1 %	0.9 %
Percent of volume	16.8 %	83.2 %

*This includes all depths.

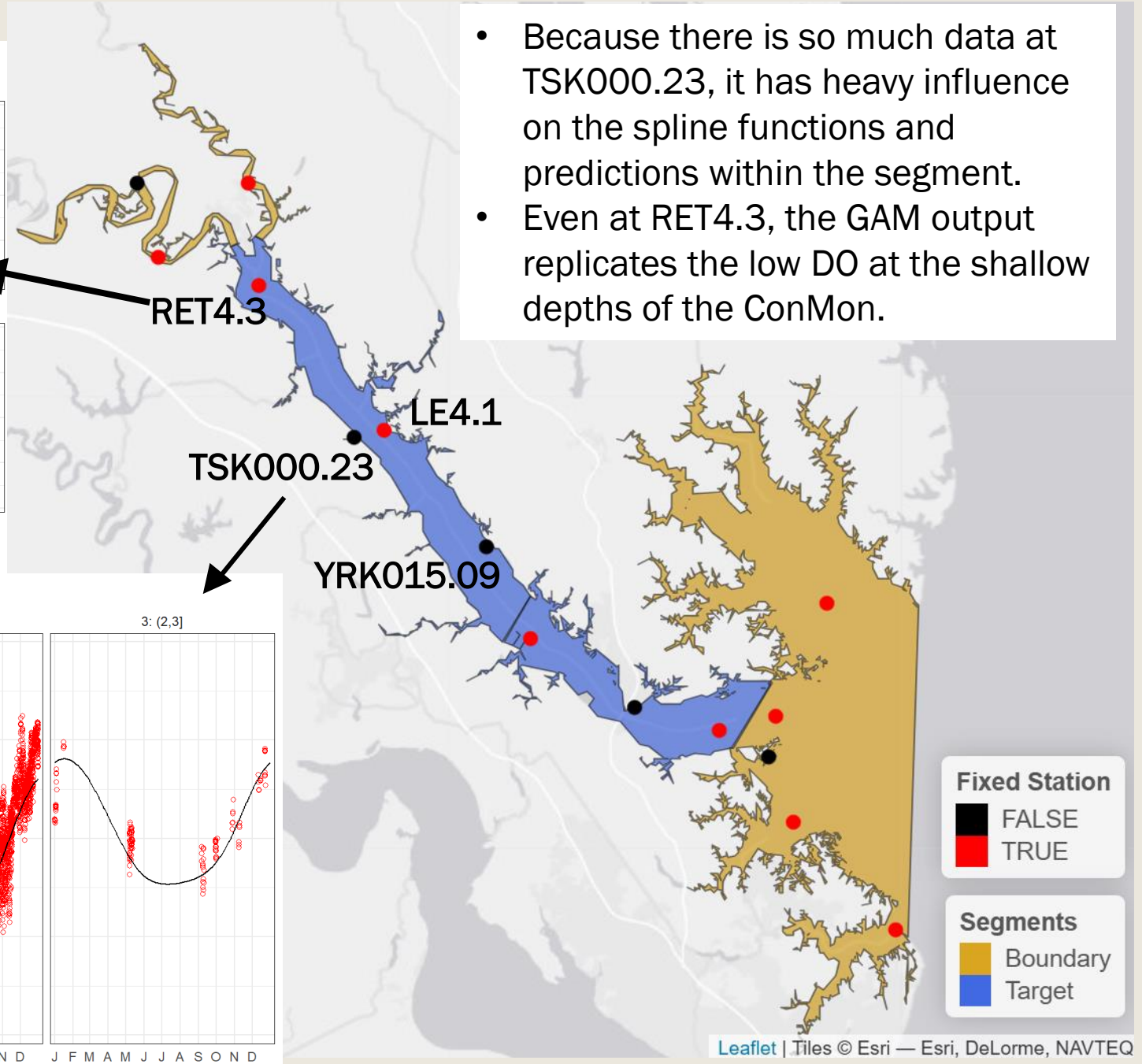


Results: no weighting





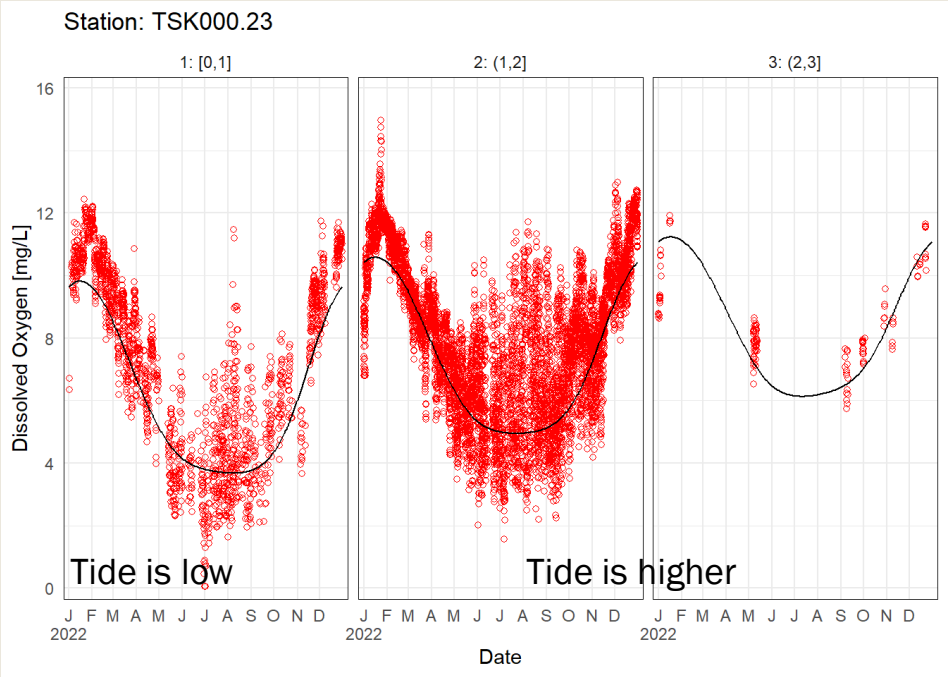
- Because there is so much data at TSK000.23, it has heavy influence on the spline functions and predictions within the segment.
- Even at RET4.3, the GAM output replicates the low DO at the shallow depths of the ConMon.



More info

- TSK000.23 is in Taskinas Creek whose depth varies from ~1 – 3 m.
- This DO pattern might be an important feature for similar creeks, but it has an out-sized impact on the spline fits throughout the York MH due to the large quantity of data.

From Google Maps

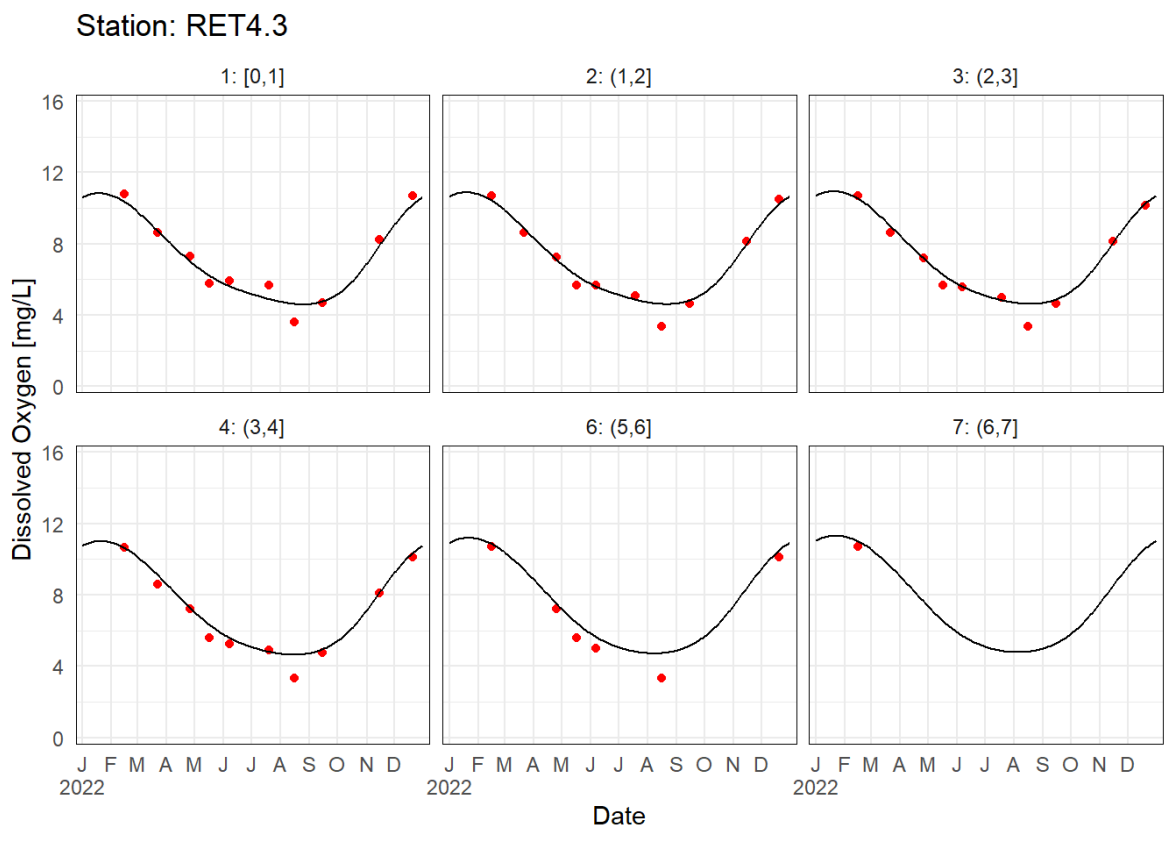


We know this situation does not happen in every segment. But looking bay-wide, there are multiple cases of high frequency stations influencing results simply due to the data count imbalance. Any single situation could be explained, but we need the interpolator to work with all data without unique solutions in each segment.

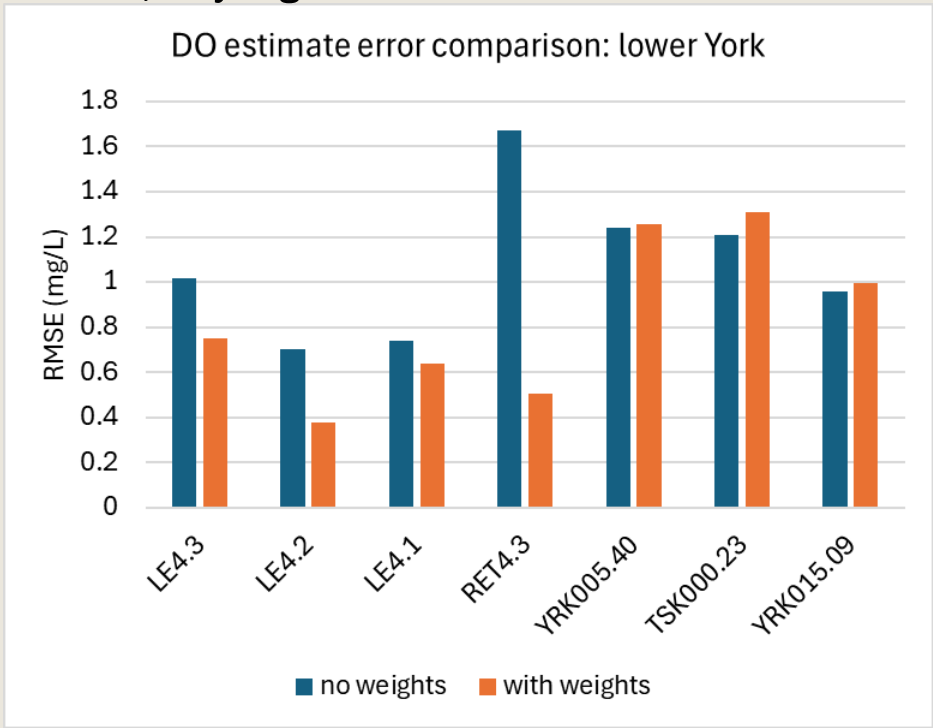
- **Approach: Weight by “Experimental Units”.**
  - *Each unique “station-date-depth|layer” is one experimental unit.*
  - *If there is >1 observation in a day, each sub-daily observation is given a smaller weight so that the weights on all samples in that day add up to 1.*
  - *This is **ONLY** for the spline fitting and does not delete any hourly data.*
- Keep in mind, for a location with the same conditions as TSK000.23 (shallow water, etc), the interpolation will heavily use that data, even with this approach.

- **Approach: Weight by "Experimental Units".**
  - *Each unique "station-date-depth|layer" is one experimental unit.*
  - *If there is >1 observation in a day, each sub-daily observation is given a smaller weight so that the weights on all samples in that day add up to 1.*
  - *This is **ONLY** for the spline fitting and does not delete any hourly data.*
- Keep in mind, for a location with the same conditions as TSK000.23 (shallow water, etc), the interpolation will heavily use that data, even with this approach.

Result at RET4.3 – Daily GAM with weighting fits through the data better



Results in the region: RMSEs decrease at fixed stations, very slight increases at ConMons



Conclusions

- We are proposing the “Experimental Unit” of a station-day-depth|layer for the mean mid-day GAM*. All hourly data will be retained, but may be weighted slightly less in fitting the GAM smooth functions.
- Comparison was done in every segment, both for 2022 and 2016.
  - *Some segments are not impacted by this misbalance of classes. It depends on the nature of the data.*
  - *When segments are impacted, the daily part of the interpolation improved across most stations with weighting.*
- This challenge is not unique to the 4-D interpolator. Likely we’d be looking at this challenge in some way if just aggregating the raw data for violations.

*Other parts of the 4-D interpolator uses **station-hour-depth** as the “Experimental Unit.”

Data Landscape

Informs mean mid-day space–time interpolation

Fixed-station network	Cruise-track monitoring (**DataFlow***)	Continuous monitoring (**ConMon***)	Vertical arrays (**NOAA**)	Additional State Agency Collected Data
✓Fixed location; broad spatial coverage ✓Long-term consistency ✓Profiles (every 1-2 meter to bottom) ✓150+ sites ✓1-2x/month	✓Surface mapping continuous data (~0.5 m) ≈ 8-10 sites/yr* ≈ 4-7 cruises per year (Apr-Oct)	✓Fixed location ✓High frequency sampling ✓Fixed depth near surface or bottom ≈ 25-30 sites/yr*; typically, 6-9 months/yr; some year round	✓Fixed location; multi-depth ✓High frequency sampling ✓New since 2022 ≈ 2-3 sites/yr ≈ 5-9 month deployments	✓Expands monitoring breadth ✓1-2x/month **Citizen monitoring (Tier 3)** ✓Expands monitoring breadth ✓1-2x/month
Vertical correlation	Horizontal correlation	Daily correlation; Daily cycle; Hourly correlation	Vertical correlation; Daily correlation; Daily cycle; Hourly correlation	

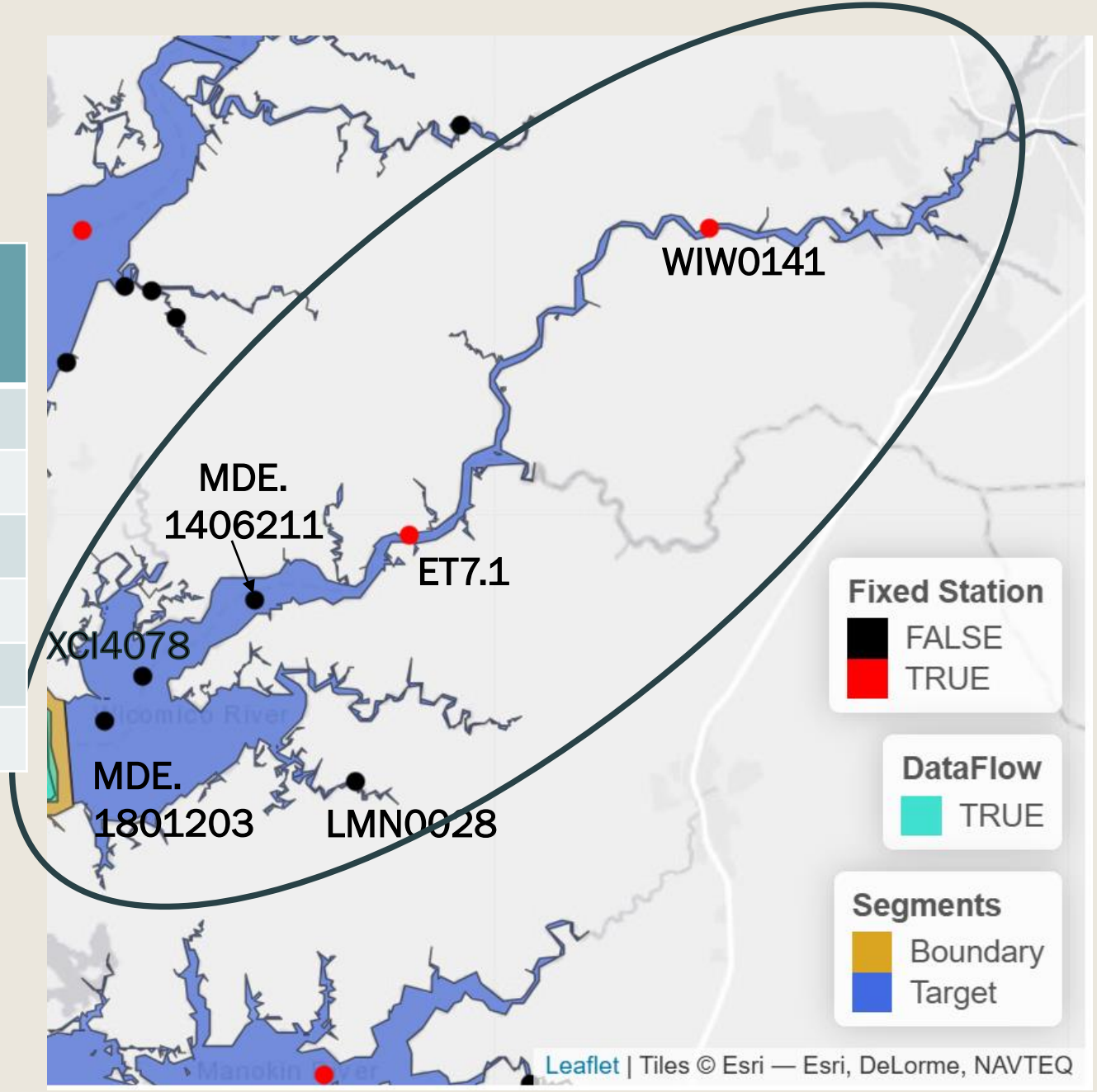
*Sites commonly rotated about every 3 years to broaden coverage

If needed: example 2

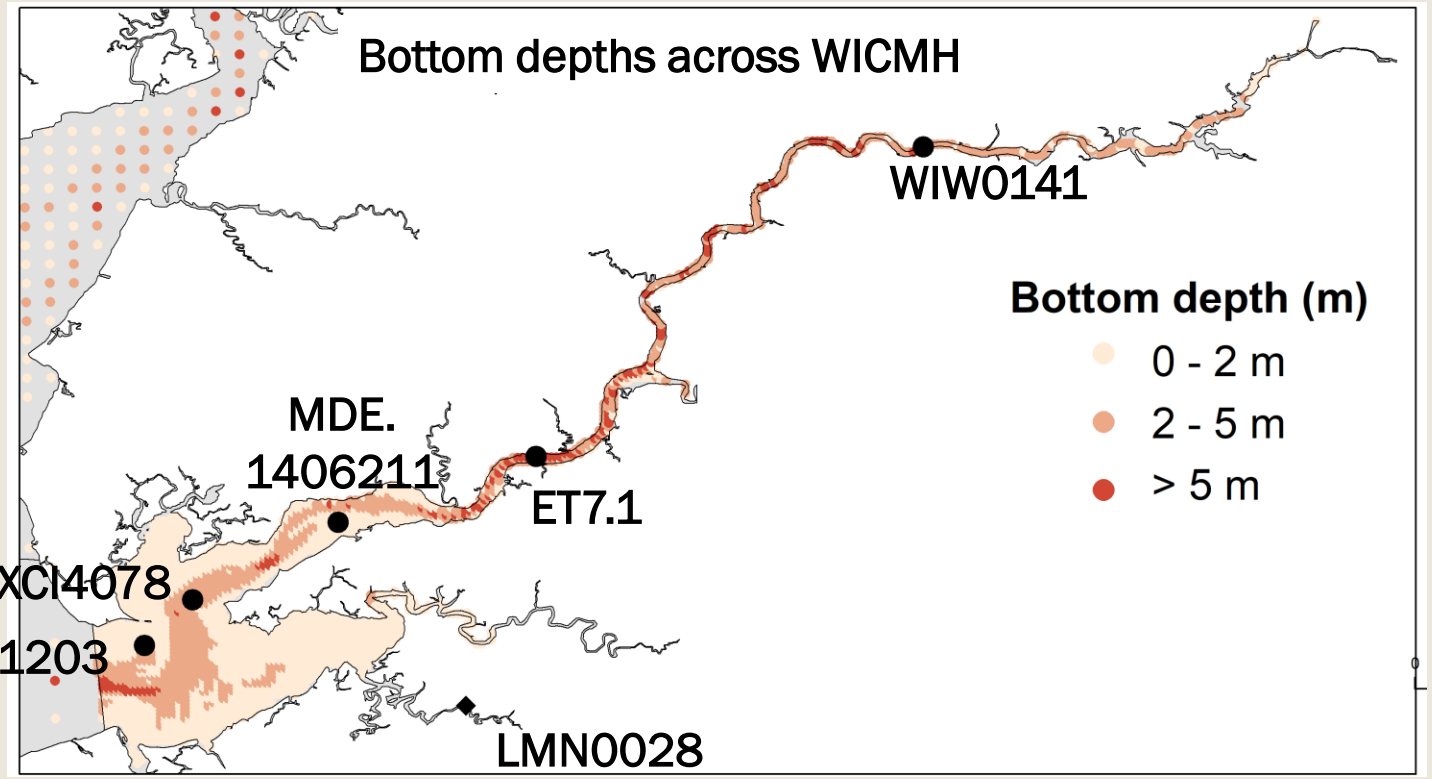
Example 2

2022 Stations in WICMH

Station	Median bottom depth (m)*	Count of samples (hour-depths)*	Count of hours with samples*
WIW0141	5.7	10 (0.13%)	10 (0.13%)
ET7.1	7.4	79 (1.0%)	12 (0.15%)
MDE.1406211	1.7	7 (0.088%)	6 (0.077%)
XCI4078	3.7	45 (0.57%)	12 (0.15%)
MDE.1801203	2.2	9 (0.11%)	5 (0.064%)
LMN0028	1.3	7,763 (98%)	7,740 (99%)



Spatial frequencies in sampling

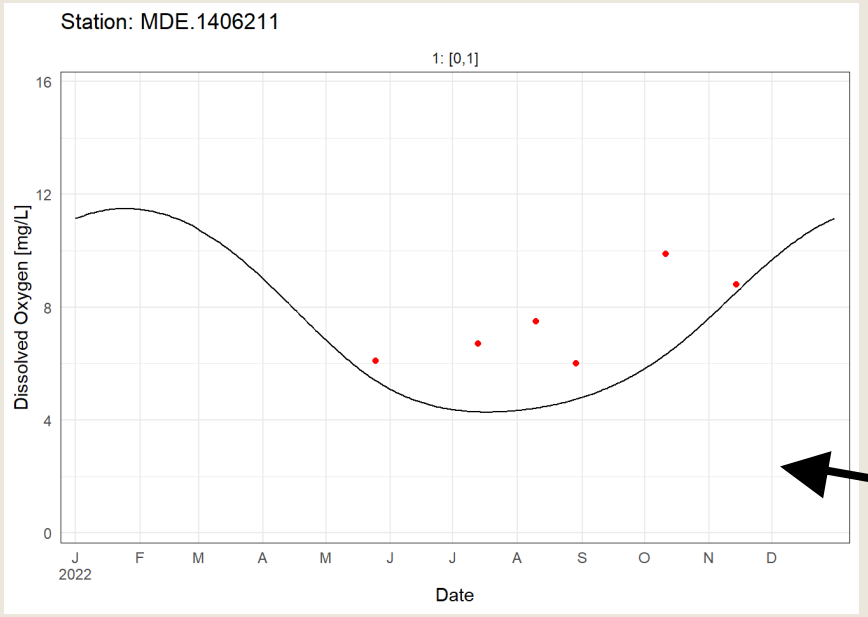


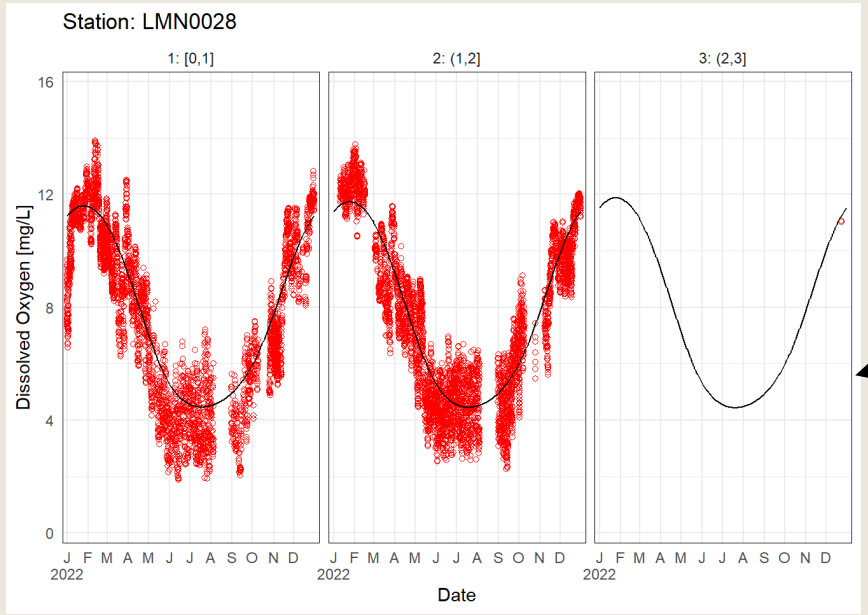
	Water 0 to 2m bottom depths	Water 2 to 12 m bottom depths
Percent of hourly samples*	98.2 %	1.8 %
Percent of volume	51.7 %	48.3 %

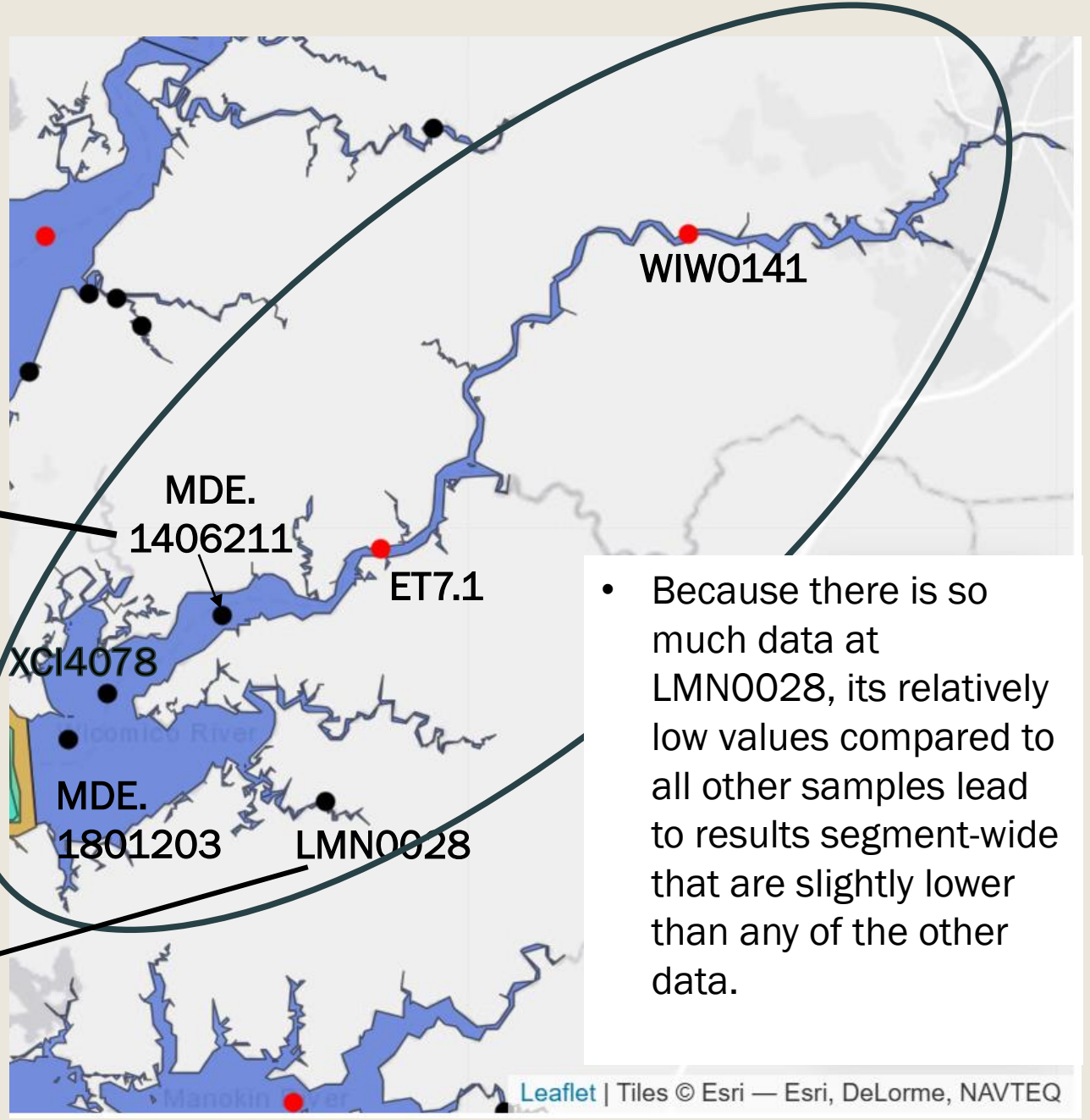
*This includes all depths.

→ Shallow waters are over-represented in this data, but not as extremely as YRKMH.

Results: no weighting



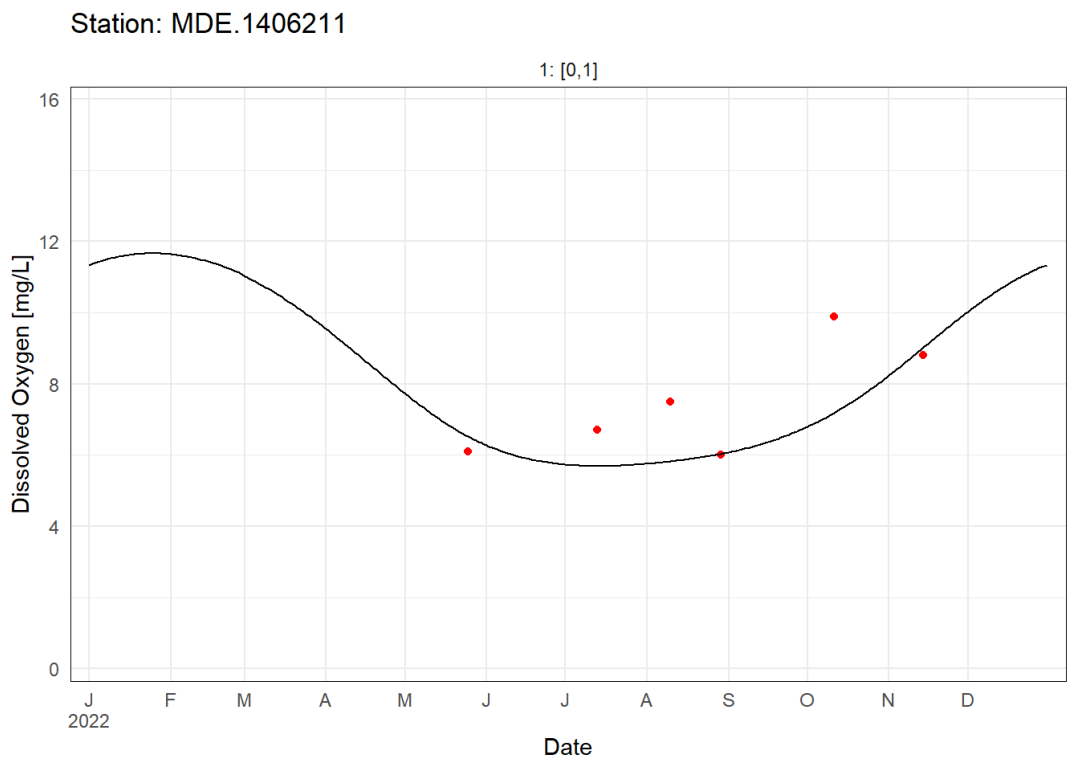




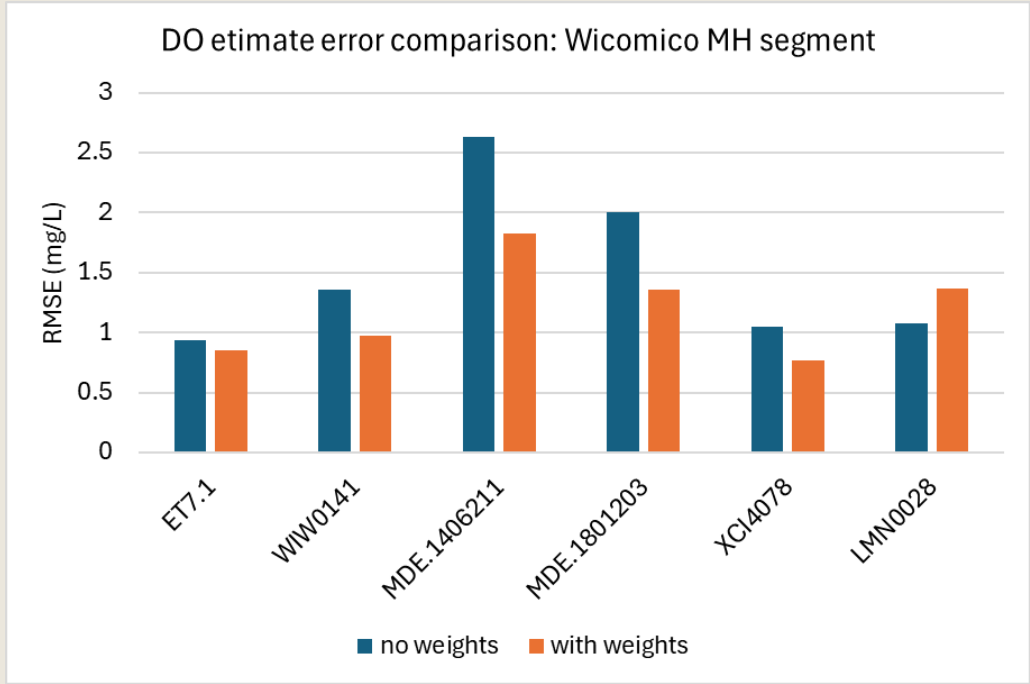
- Because there is so much data at LMN0028, its relatively low values compared to all other samples lead to results segment-wide that are slightly lower than any of the other data.

- Approach: Weight by “Experimental Units”.
  - *Each unique “station-date-depth|layer” is one experimental unit.*
  - *If there is >1 observation in a day, each sub-daily observation is given a smaller weight so that the weights on all samples in that day add up to 1.*
  - *This is **ONLY** for the spline fitting and does not delete any hourly data.*
- Keep in mind, for a location with the same conditions as LMN0028 (shallow water, etc), the interpolation will heavily use that data, even with this approach.

Result at MDE.1406211 – Daily GAM with weighting fits min better



Results in the region: RMSEs decrease at fixed stations, increases at ConMon



extras

Temporal frequencies in sampling

- It is a huge success that these ConMon stations cover almost every hour of the year!
  - *That teaches us a lot about shallow water DO year-round and we need this information in the 4-D interpolator.*
- However, if the ConMon data are not representative of the deeper waters, we want to be sure they impact the interpolation in only the right places and times.

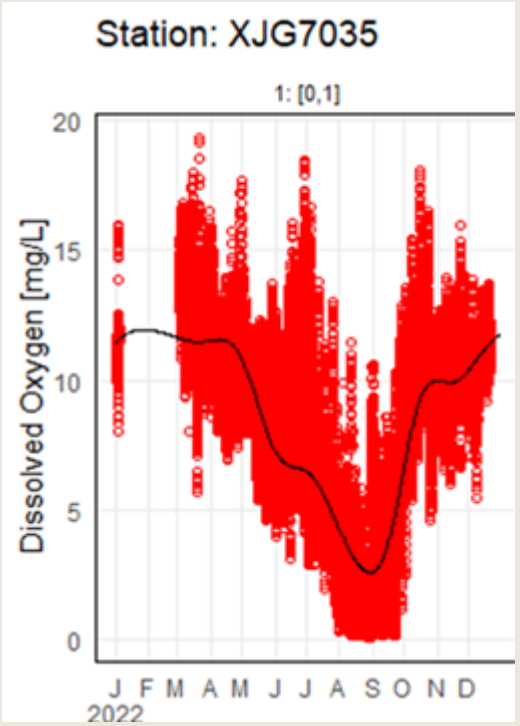
For YRKMH data in 2022

	One or both ConMon stations	Sampled by either Fixed station
Count of hours with samples	8,724 hours	20 hours
Percent of hours/year	99.6 %	0.228 %

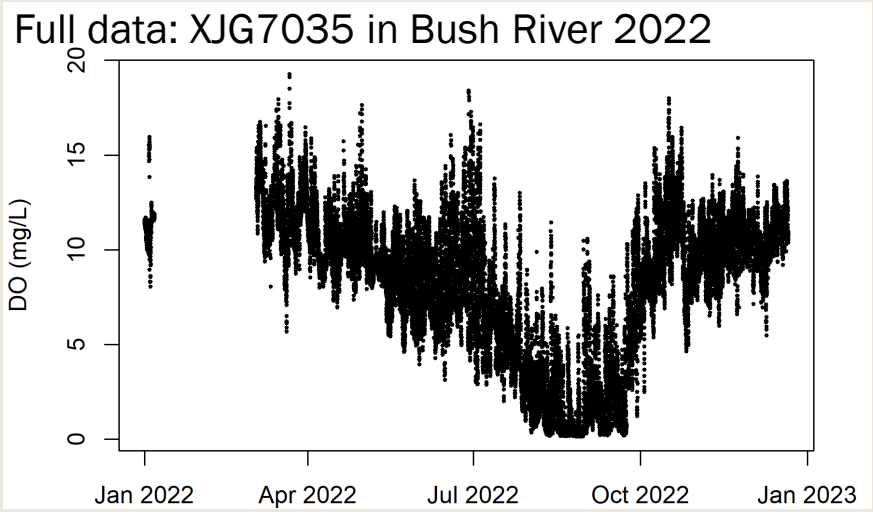
For this part only

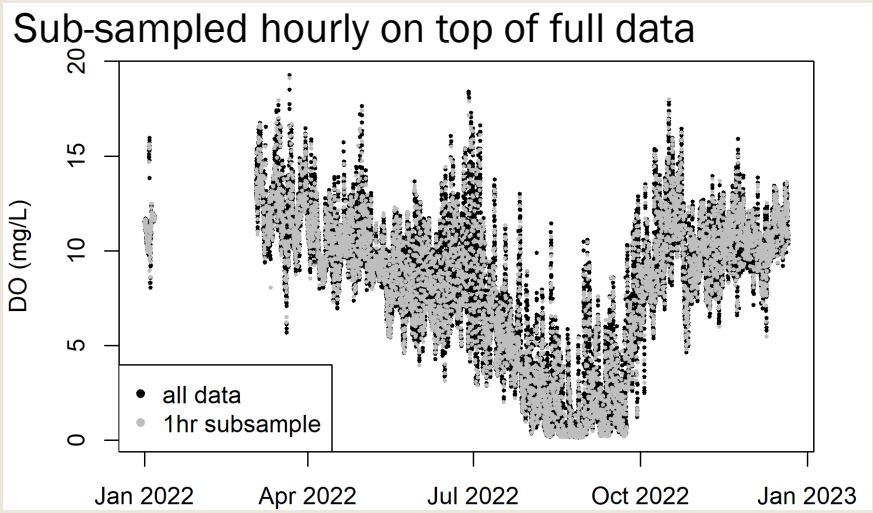
Mean mid-day space-and-time interpolation

What if ConMon is sub-sampled to hourly?

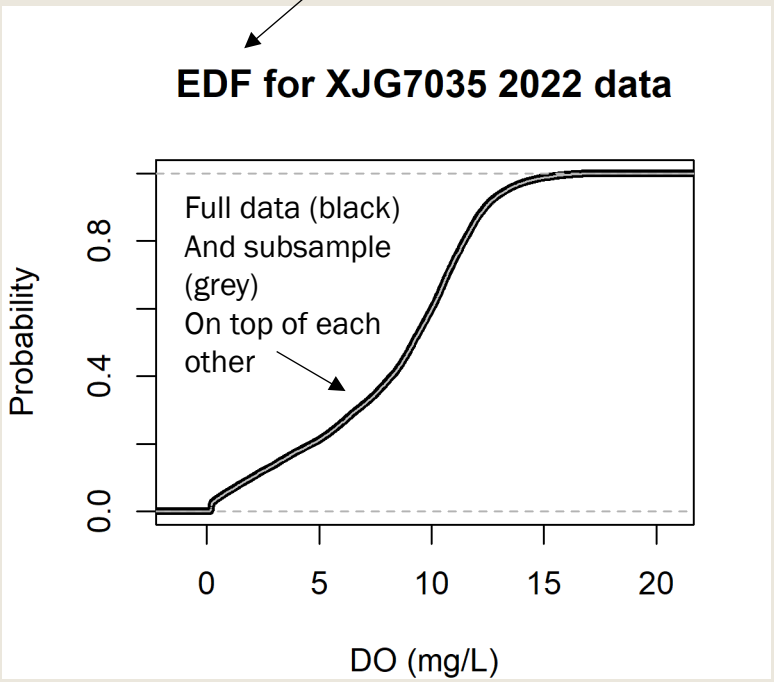


=





EDF = Empirical density function



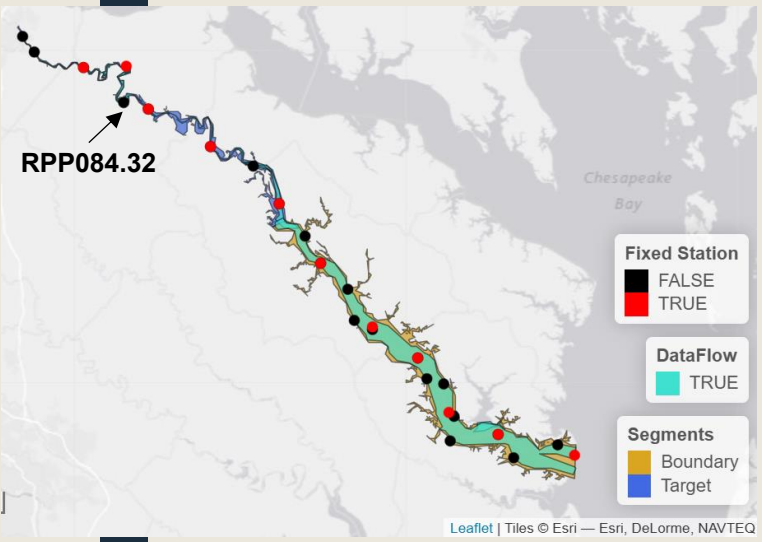
Data set	count	10th percentile	Fraction <3.2 mg/L	Fraction < 5mg/L
All	28,127	1.97	0.147	0.212
Sub-sample	7,071	1.98	0.146	0.211

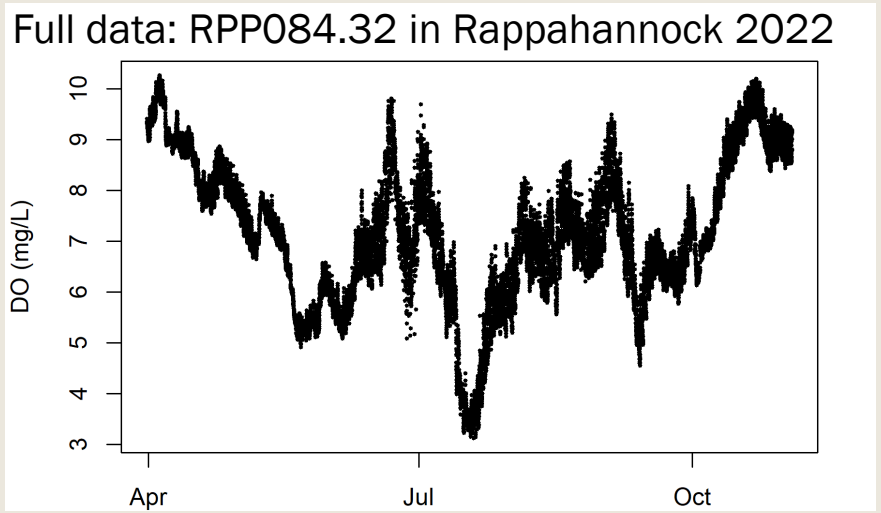
These summaries suggest we are not changing the important features of this dataset by sub-sampling this 15 min data to 1 hour.

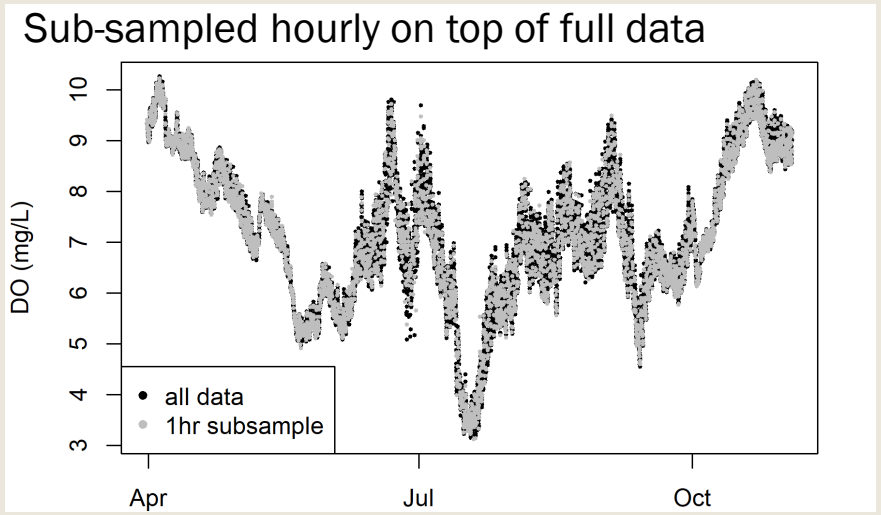
For this part only

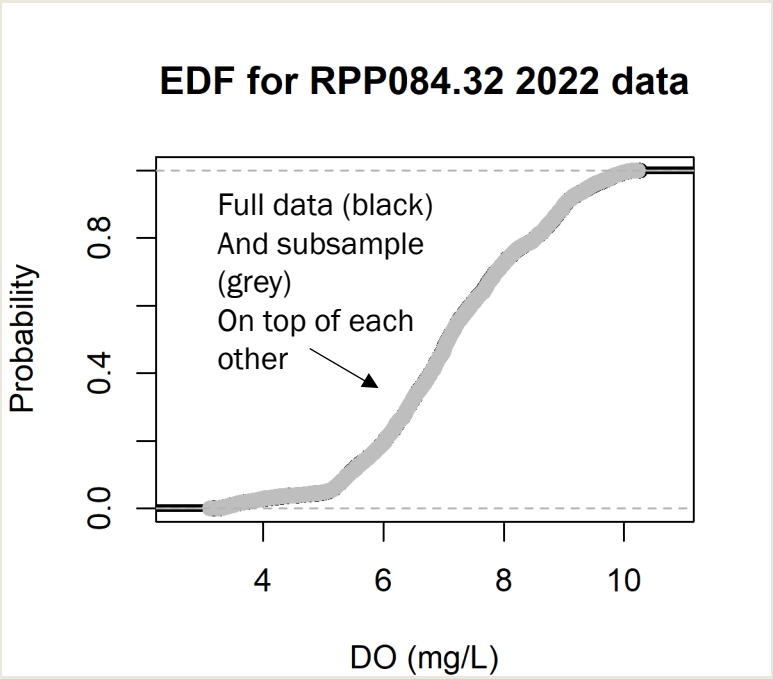
Mean mid-day space-and-time interpolation

Another example









Data set	count	10th percentile	Fraction <3.2 mg/L	Fraction < 5mg/L
All	20,927	5.42	0.00029	0.0467
Sub-sample	5,316	5.42	0.00038	0.0478

These summaries suggest we are not changing the important features of this dataset by sub-sampling this 15 min data to 1 hour.

Purpose: Build a tool for more complete criteria assessment

DO criteria that currently can be evaluated with existing approaches and data

Table 1. Chesapeake Bay dissolved oxygen criteria.

Designated Use	Criteria Concentration/Duration	Protection Provided	Temporal Application
Migratory fish spawning and nursery use *	7-day mean ≥ 6 mg liter^{-1} (tidal habitats with 0-0.5 ppt salinity)	Survival/growth of larval/juvenile tidal-fresh resident fish; protective of threatened/endangered species.	February 1 - May 31
	Instantaneous minimum ≥ 5 mg liter^{-1}	Survival and growth of larval/juvenile migratory fish; protective of threatened/endangered species.	
	Open-water fish and shellfish designated use criteria apply		June 1 - January 31
Shallow-water bay grass use	Open-water fish and shellfish designated use criteria apply		Year-round
Open-water fish and shellfish use	30-day mean ≥ 5.5 mg liter^{-1} (tidal habitats with 0-0.5 ppt salinity)	Growth of tidal-fresh juvenile and adult fish; protective of threatened/endangered species.	Year-round
	30-day mean ≥ 5 mg liter^{-1} (tidal habitats with >0.5 ppt salinity)	Growth of larval, juvenile and adult fish and shellfish; protective of threatened/endangered species.	
	7-day mean ≥ 4 mg liter^{-1}	Survival of open-water fish larvae.	
	Instantaneous minimum ≥ 3.2 mg liter^{-1}	Survival of threatened/endangered sturgeon species.[1]	
Deep-water seasonal fish and shellfish use	30-day mean ≥ 3 mg liter^{-1}	Survival and recruitment of bay anchovy eggs and larvae.	June 1 - September 30
	1-day mean ≥ 2.3 mg liter^{-1}	Survival of open-water juvenile and adult fish.	
	Instantaneous minimum ≥ 1.7 mg liter^{-1}	Survival of bay anchovy eggs and larvae.	
	Open-water fish and shellfish designated-use criteria apply		October 1 - May 31
Deep-channel seasonal refuge use	Instantaneous minimum ≥ 1 mg liter^{-1}	Survival of bottom-dwelling worms and clams.	June 1 - September 30
	Open-water fish and shellfish designated use criteria apply		October 1 - May 31

[1] At temperatures considered stressful to shortnose sturgeon (>29°C), dissolved oxygen concentrations above an instantaneous minimum of 4.3 mg liter^{-1} will protect survival of this listed sturgeon species.

*Note a 30-day mean 6 mg/L MSN value is evaluated for purpose of the WQ indicator.

From EPA 2003 Ambient Water Quality Criteria