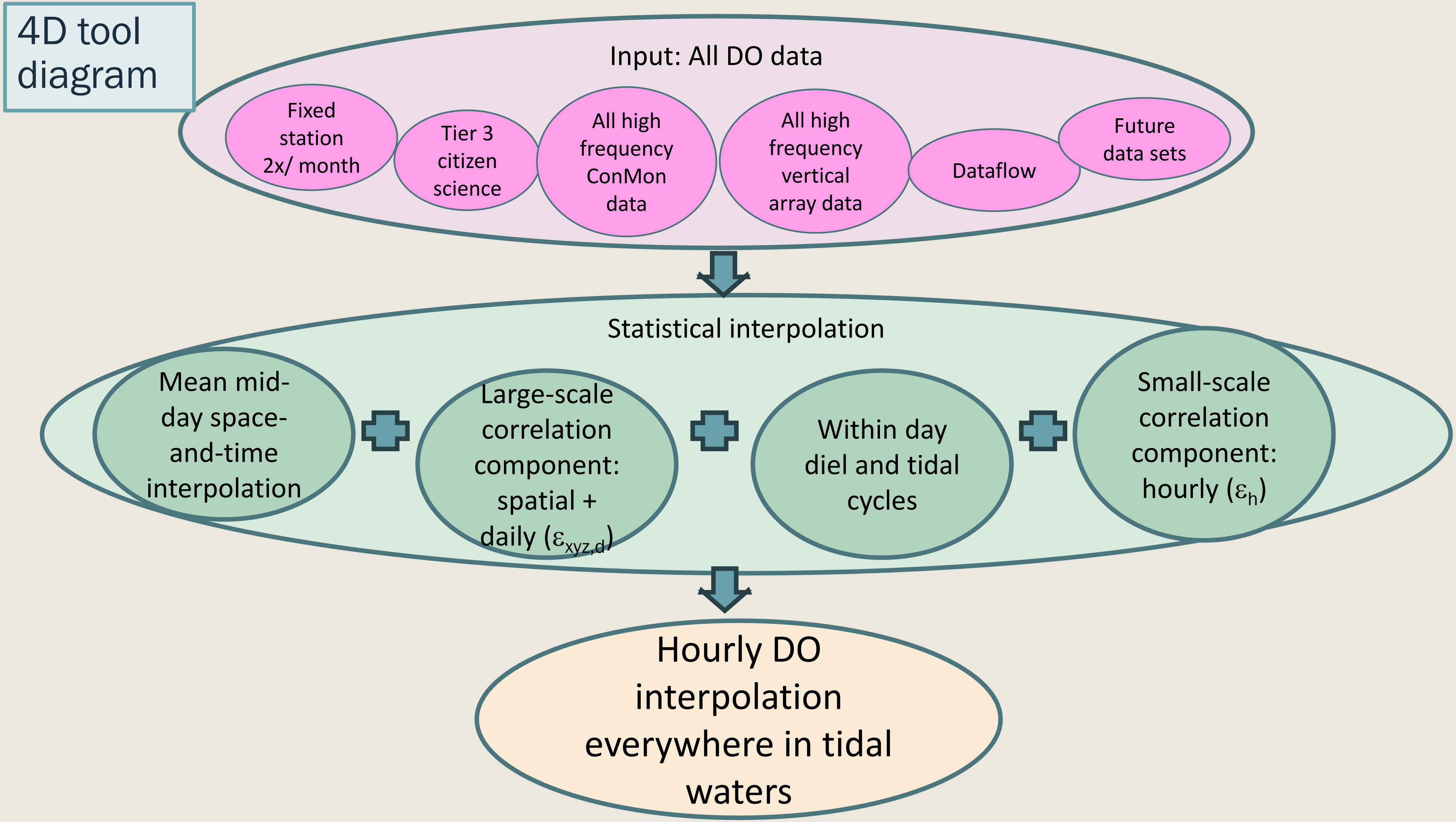
Sub-sampling Continuous Monitoring and Dataflow data

BORG meeting
Nov. 17, 2025

Rebecca Murphy (UMCES/CBP), Jon Harcum (Tetra Tech),

Elgin Perry (statistics consultant)

4D tool diagram



Input: All DO data

Fixed station 2x/ month

Tier 3 citizen science

All high frequency ConMon data

All high frequency vertical array data

Dataflow

Future data sets

Statistical interpolation

Mean mid-day space-and-time interpolation

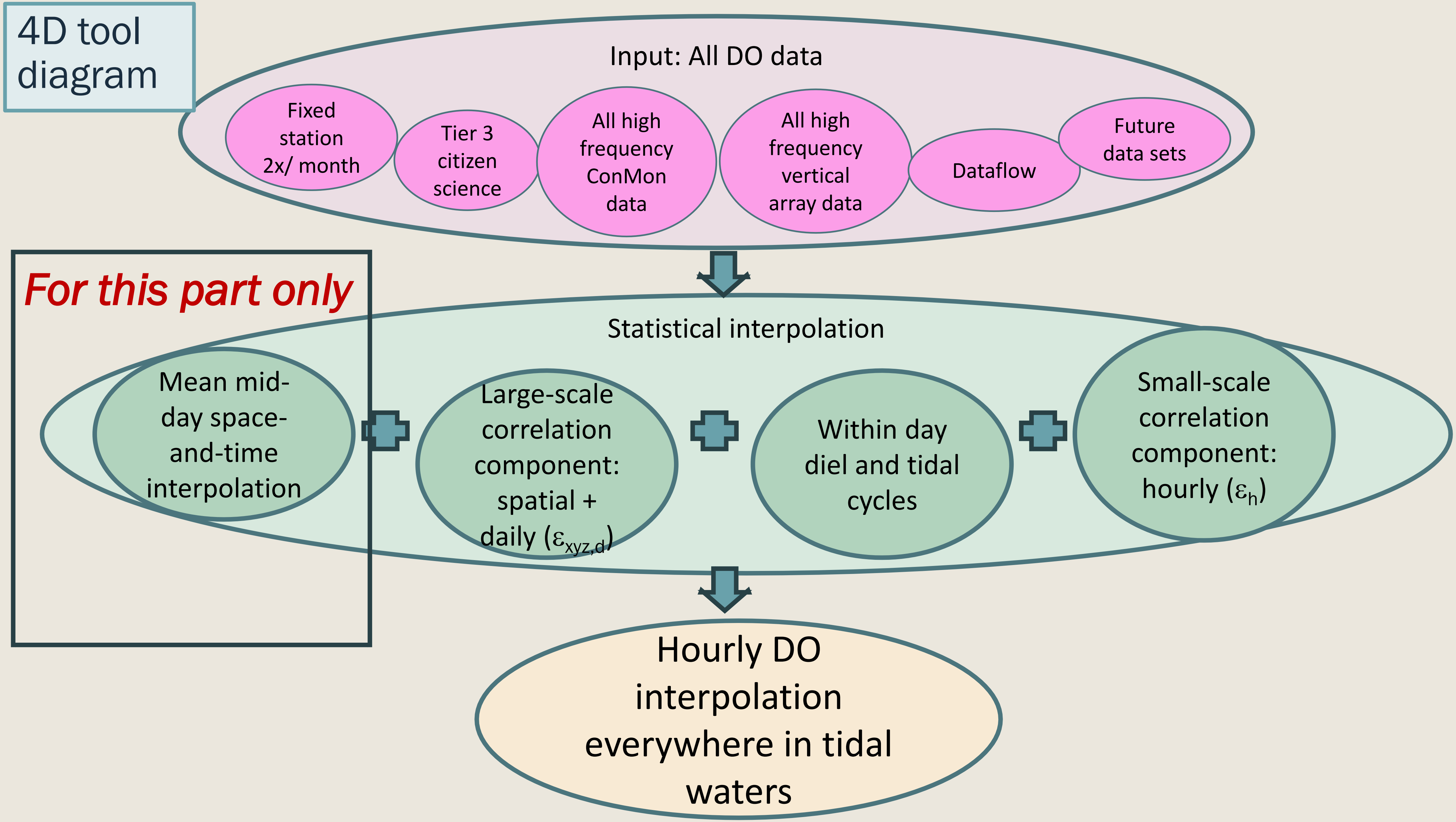
Large-scale correlation component: spatial + daily ($\varepsilon_{xyz,d}$)

Within day diel and tidal cycles

Small-scale correlation component: hourly (ε_h)

Hourly DO interpolation everywhere in tidal waters

4D tool diagram



Input: All DO data

Fixed station 2x/ month

Tier 3 citizen science

All high frequency ConMon data

All high frequency vertical array data

Dataflow

Future data sets

For this part only

Statistical interpolation

Mean mid-day space-and-time interpolation

Large-scale correlation component: spatial + daily ($\varepsilon_{xyz,d}$)

Within day diel and tidal cycles

Small-scale correlation component: hourly (ε_h)

Hourly DO interpolation everywhere in tidal waters

For this part only

Mean mid-day space-and-time interpolation

Context

- **This summer, we updated the temporal frequency of the high frequency data used in the mid-day interpolation based on feedback from this group:**
  - → *Previously we used a daily sub-sample of any high frequency data (since we are predicting daily).*
  - → *Feedback indicated the team wanted all high frequency data input to this part of the tool.*
  - → *We pivoted and tried to fit the mean mid-day interpolation with all available high-frequency data. This includes readings taken every 10 or 15 minutes from ConMon stations, as well as all Dataflow inputs.*

For this part only

Mean mid-day space-and-time interpolation

Context

- What we saw:

  → *Sometimes, DO at a single ConMon station behaves differently from other DO within its region. When this happens, the large amount of data at that ConMon can disproportionately affect the interpolation results for the whole region.*

  → *This is not unprecedented. Elgin is doing some literature review, and this is called "imbalance of classes" in the data science literature when data representing different conditions (shallow vs. deep in our case) are sampled at different frequencies.*

For this part only

Mean mid-day space-and-time interpolation

Context

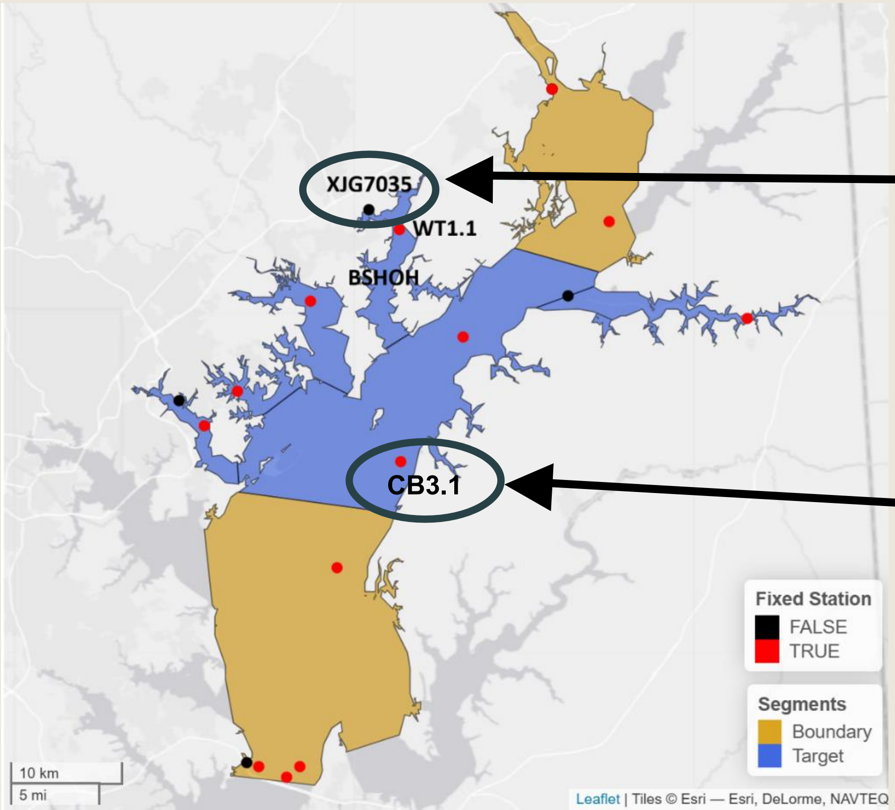
- **We’re considering two approaches to balance the call to use all data vs. challenge of “imbalance of classes”:**
  1. ***Use hourly frequency:*** *We propose to sub-sample the high-frequency data so it matches the resolution of the final output—about one reading per hour for ConMon and spaced roughly 500 meters for Dataflow.*
  2. ***Update the smoothing method:*** *We may also need to adjust the daily smoothing functions and their use of the different types of data “classes” to better handle the higher-frequency data.*

For this part only

Mean mid-day space-and-time interpolation

Example of high frequency impact

CB2OH including Bush River and other tribs



Long term ConMon.
In 2022: 28,127 DO samples

Fixed station.
In 2022: 208 DO samples

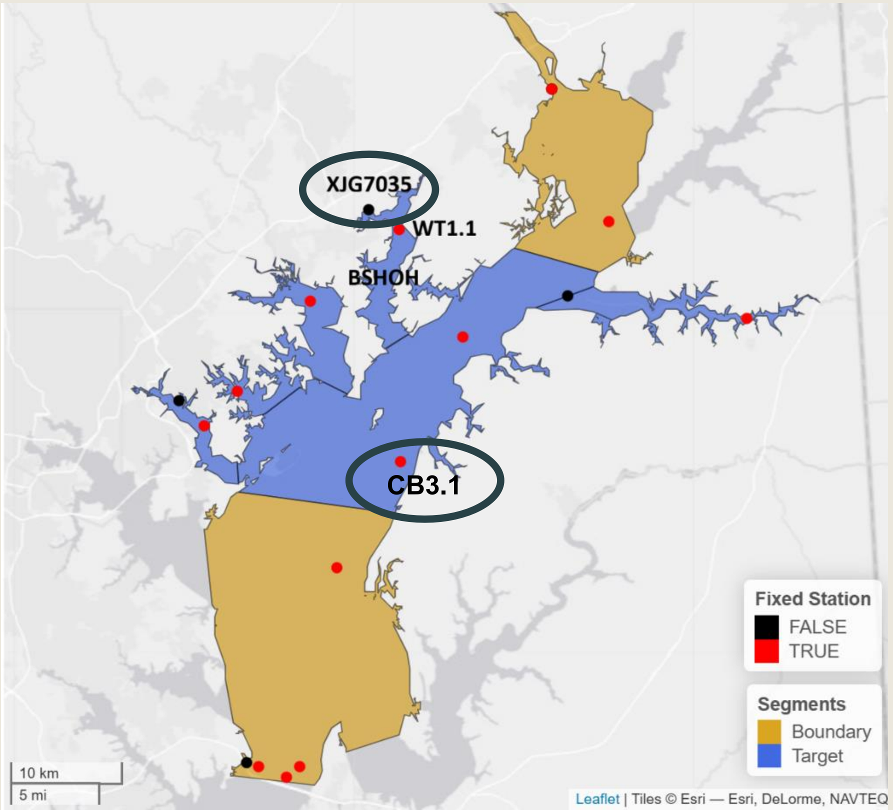
98.8% of the DO in these grouped segments is from 2 ConMon in 2022.

For this part only

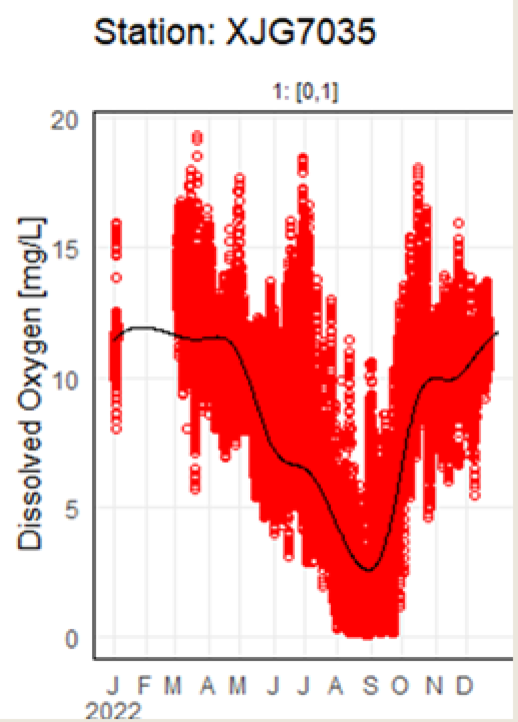
Mean mid-day space-and-time interpolation

Example of high frequency impact

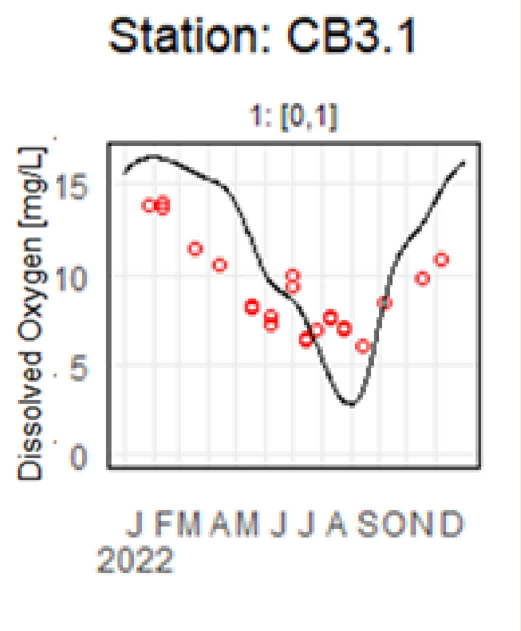
CB2OH including Bush River and other tribs



Example: fairly low DO in Aug-Oct observed at the ConMon relative to other surface observations



Result: Too much influence on surface interpolation results, even far away from this station



For this part only

Mean mid-day space-and-time interpolation

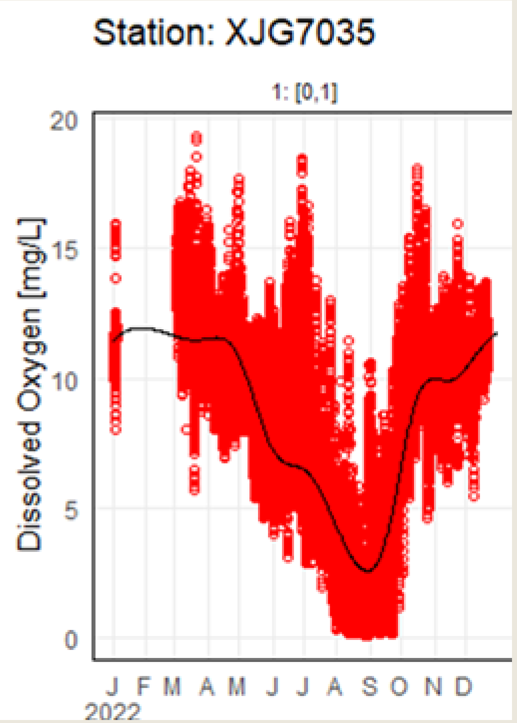
What if ConMon is sub-sampled to hourly?

- It looks like it helps represent the spatial as well as temporal patterns of each station better in the results *(not shown due to work-in-progress).*
- However, first we want to check that we aren’t changing the conclusions about the distribution of the high frequency data.

For this part only

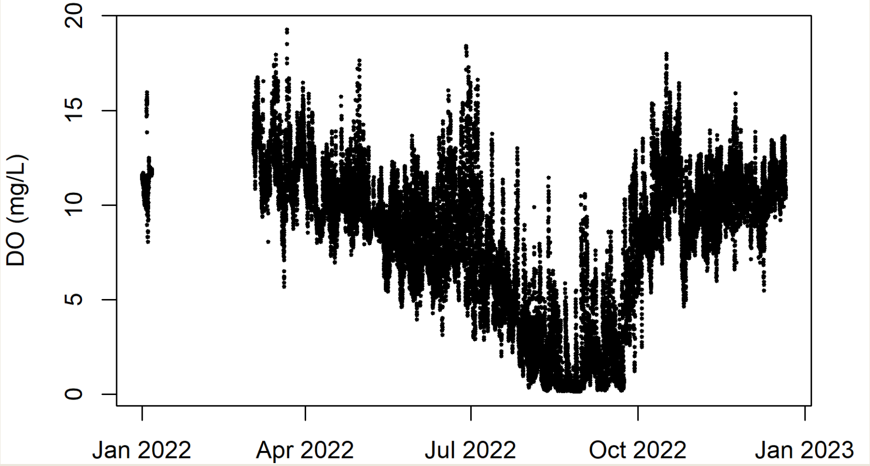
Mean mid-day space-and-time interpolation

What if ConMon is sub-sampled to hourly?

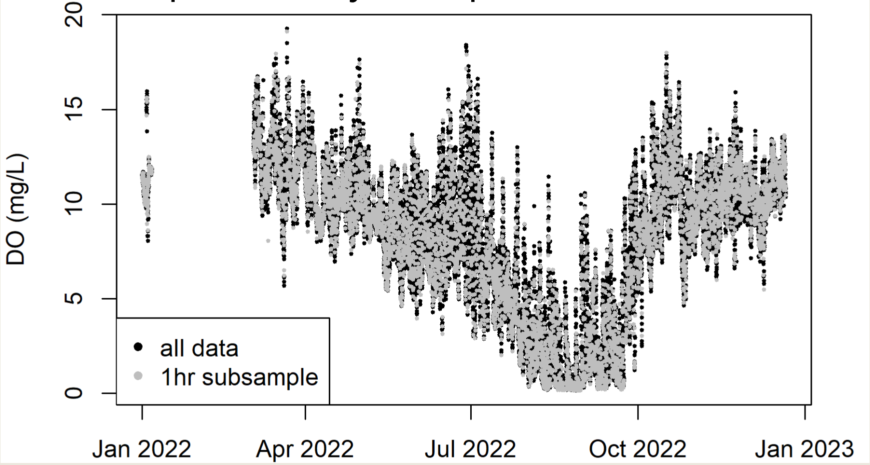


=

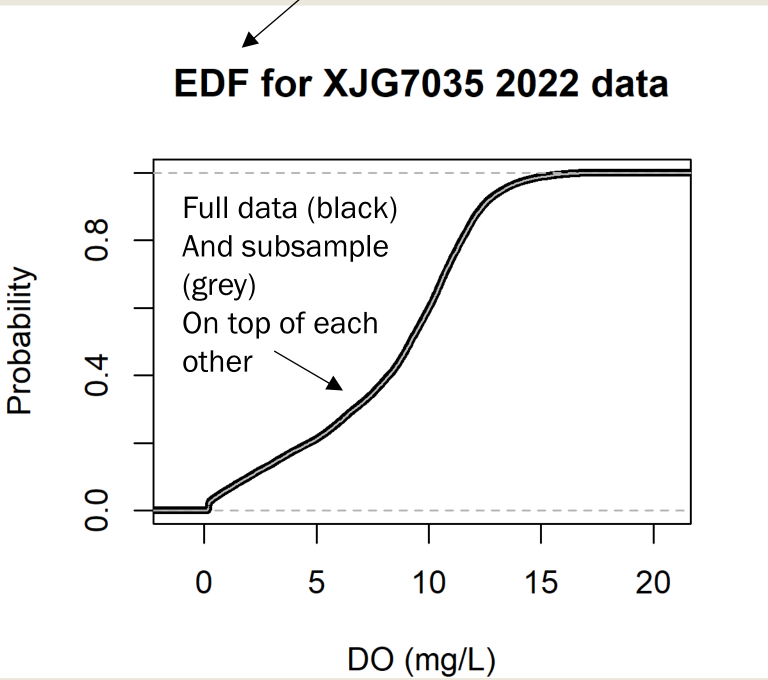
Full data: XJG7035 in Bush River 2022



Sub-sampled hourly on top of full data



EDF = Empirical density function



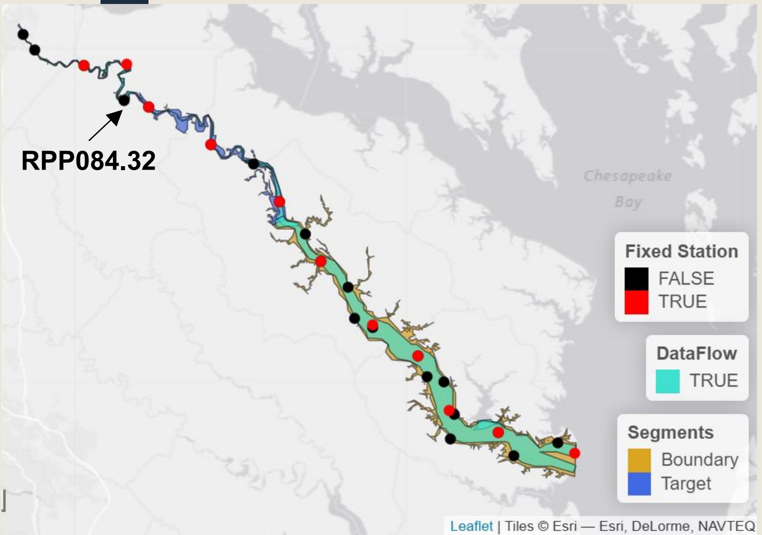
Data set	count	10th percentile	Fraction <3.2 mg/L	Fraction < 5mg/L
All	28,127	1.97	0.147	0.212
Sub-sample	7,071	1.98	0.146	0.211

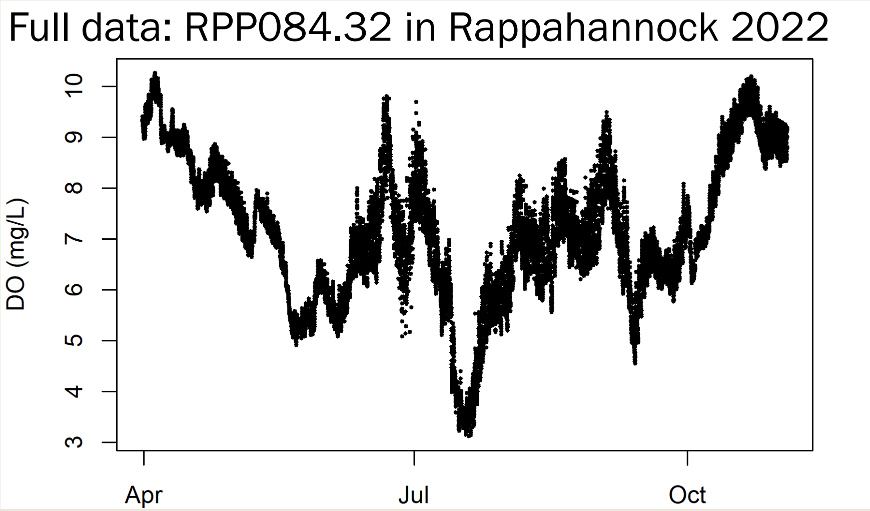
These summaries suggest we are not changing the important features of this dataset by sub-sampling this 15 min data to 1 hour.

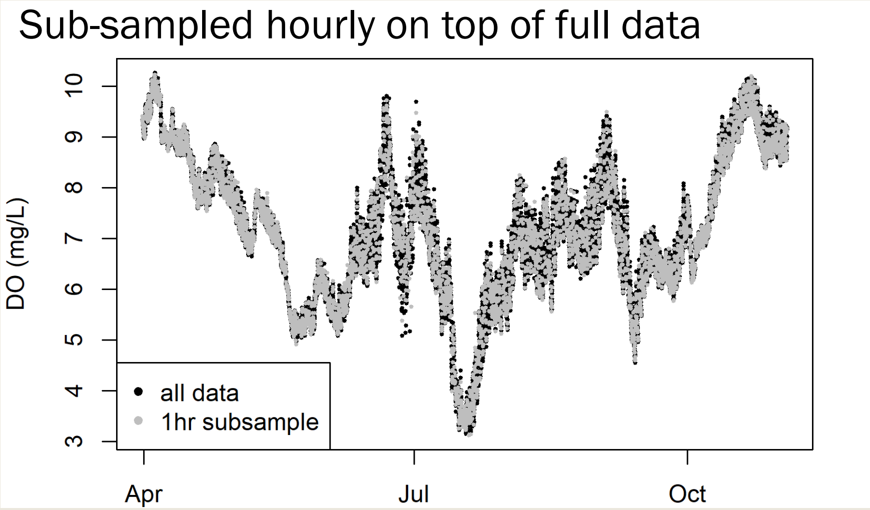
For this part only

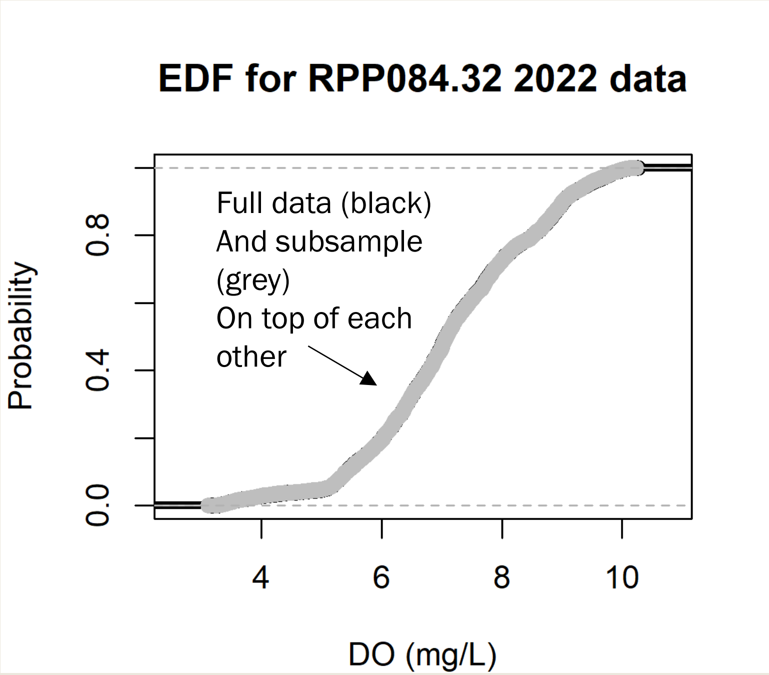
Mean mid-day space-and-time interpolation

Another example




Full data: RPP084.32 in Rappahannock 2022


Sub-sampled hourly on top of full data


EDF for RPP084.32 2022 data

Full data (black)
And subsample
(grey)
On top of each
other

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.

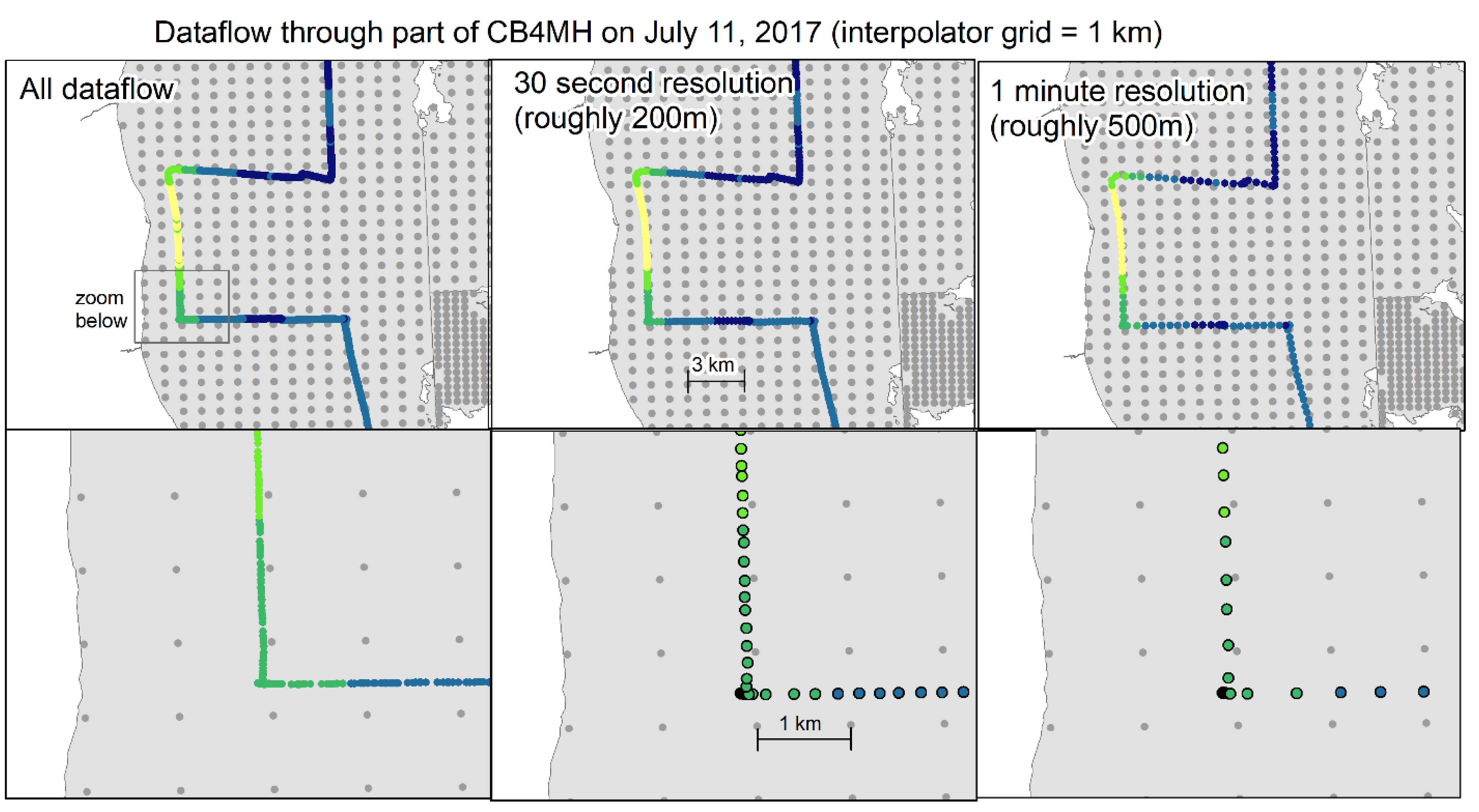
For this part only

Mean mid-day space-and-time interpolation

Dataflow subsampling

- Dataflow is different from ConMon in that the location is changing as high frequency data is collected.
- To subsample the Dataflow, we are suggesting to retain a sample every 500m or 1km. This for consistency with the interpolator grid throughout most of the bay.
- Testing is needed on the exact distance to use.
- Currently we tested sub-sampling in time, as a proxy for distance.

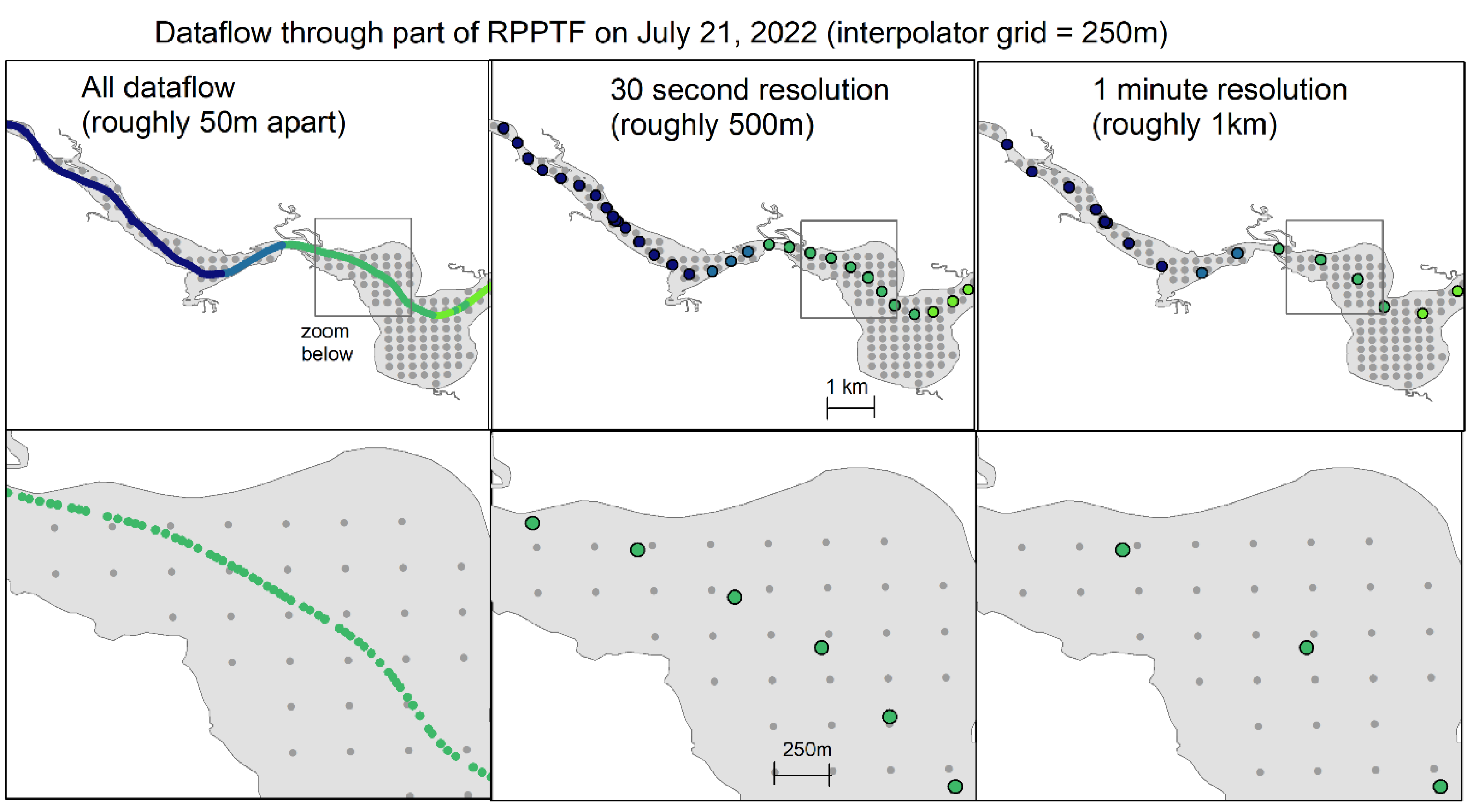
Dataflow: very dense spatial resolution



Dataflow through part of CB4MH on July 11, 2017 (interpolator grid = 1 km)

All dataflow
zoom below
30 second resolution (roughly 200m)
3 km
1 minute resolution (roughly 500m)
1 km

Dataflow: very dense spatial resolution

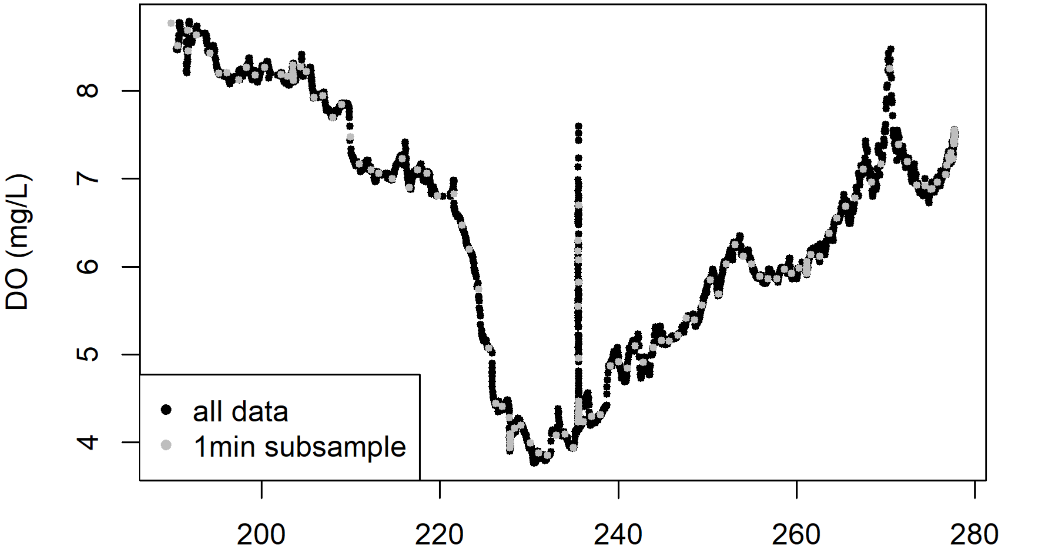


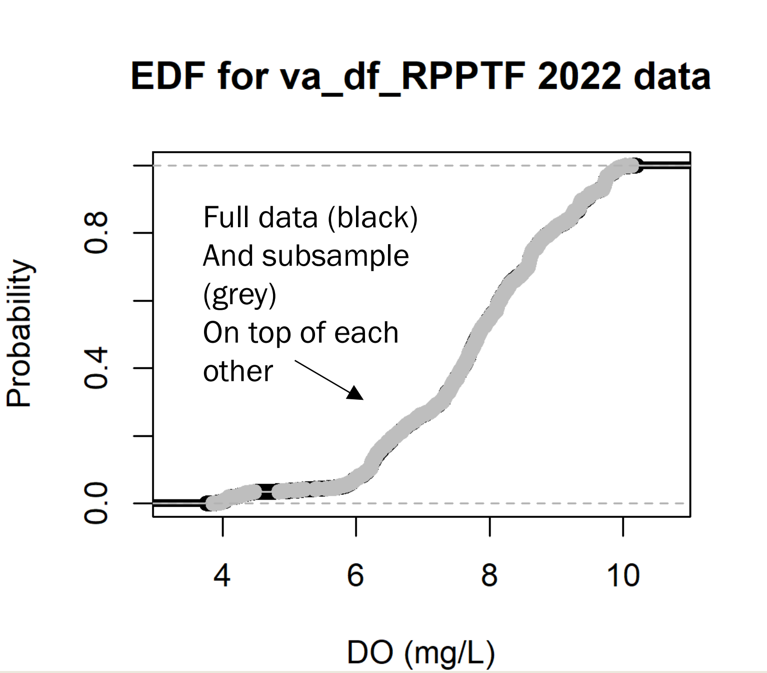
For this part only

Mean mid-day space-and-time interpolation

Subsampling impact: RPPTF dataflow in 2022

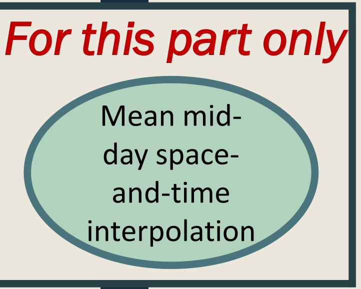
July 21, 2022: Sub-sampled DF on top of full data





Data set	count	10th percentile	Fraction <3.2 mg/L	Fraction < 5mg/L
All	29,295	6.20	0	0.0374
Sub-sample	1,423	6.19	0	0.0379

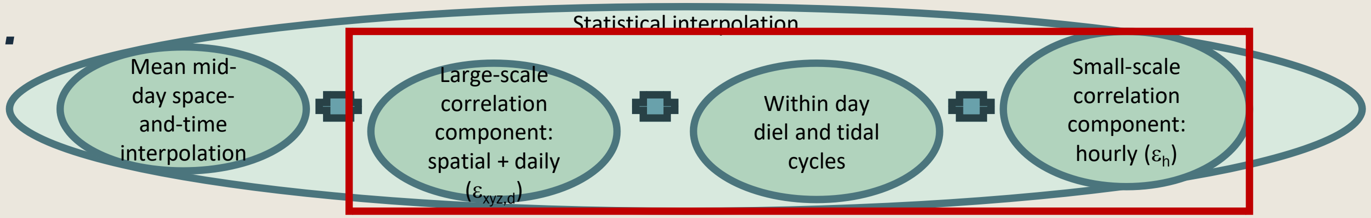
Similar conclusion for sub-sampling dataflow. However, it is likely we will subsample based on distance instead of time in the final application.



Next steps

- We will include high frequency data at the hourly level (and dataflow at ~500m) for the mid-day interpolation.
- Elgin will continue a literature review to glean how this challenge has been dealt with in related fields.
- As we move into case studies, we may need to implement additional adjustments on how the smooth functions deal with data at different frequencies to ensure we still get an appropriate spatial picture when high frequency data is included in a segment.

→ *But keep in mind, this is all for the daily estimate, and we will still add daily cycles and high frequency variability on top of any daily interpolation.*


Statistical interpolation
Mean mid-day space-and-time interpolation
Large-scale correlation component: spatial + daily ($\varepsilon_{xyz,d}$)
Within day diel and tidal cycles
Small-scale correlation component: hourly (ε_h)

extra

Purpose: Build a tool for more complete criteria assessment

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

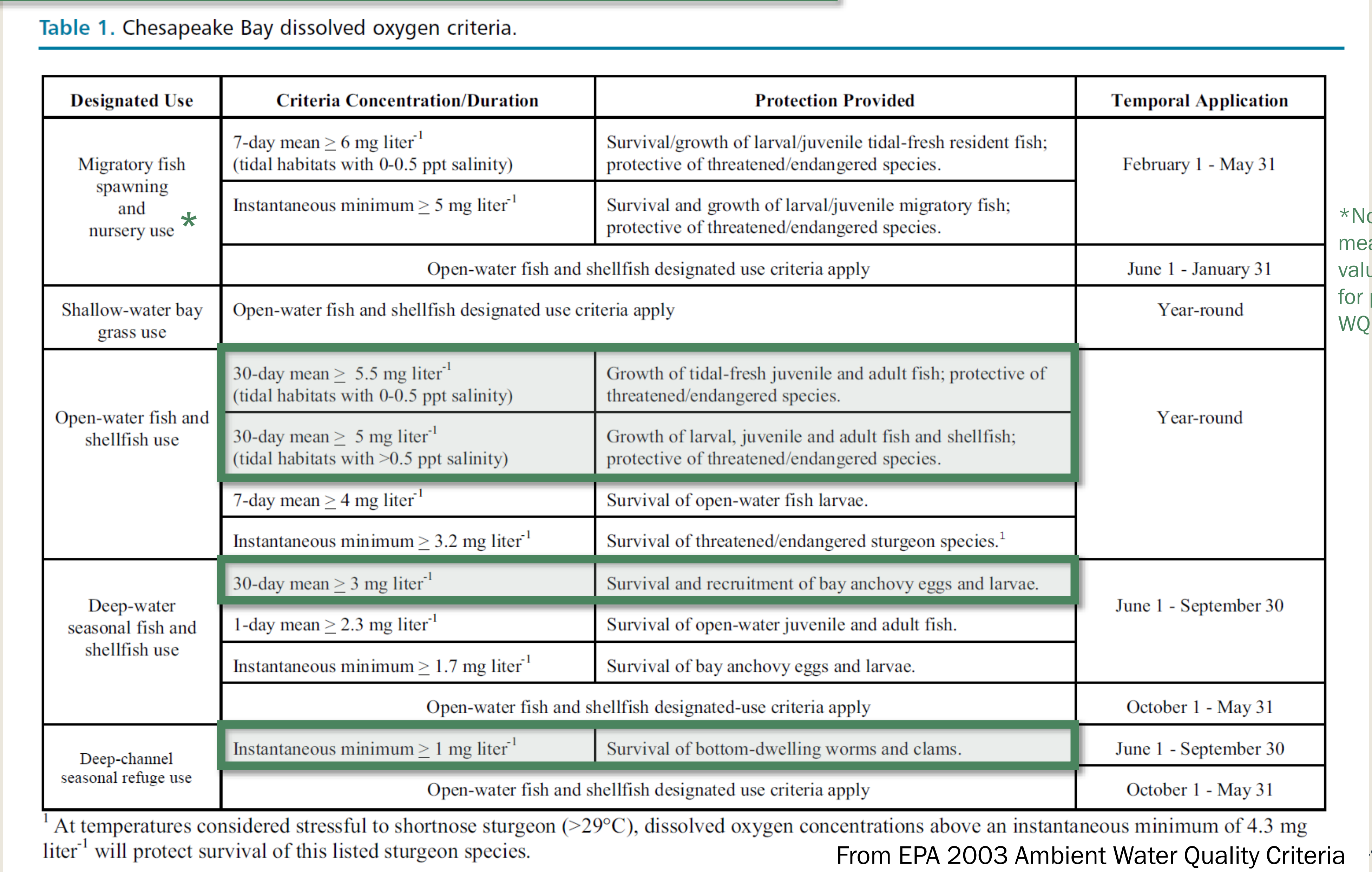


Table 1. Chesapeake Bay dissolved oxygen criteria.

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