

Testing Segment Interpolation Regions for GAM portion of 4-D interpolator

July 8, 2025

Rebecca Murphy (UMCES/CBP)

with input from Jon Harcum (Tetra Tech), Elgin Perry (consultant), Breck Sullivan, and Peter Tango (USGS)

Introduction

For the 4D spatial-temporal interpolator being developed for use in Chesapeake Bay tidal waters, one part of the interpolation process involves fitting Generalized Additive Models (GAMs) to dissolved oxygen (DO) concentrations over time by region. Those GAMs are then used to generate estimates of mid-day DO throughout each of their applicable grids, filling in the gaps between data in space and time. Details of the GAMs have been presented elsewhere, but in brief, the GAM process involves knowing the location, depth, time, and day of each DO observation; fitting smoothly varying functions between DO and these variables; and then using those fitted relationships and DO data to estimate DO in the places and times where there is no data. The GAM is only one part of the 4D interpolation. Its estimates will be added to: 1) hourly DO interpolations of the high frequency data, and 2) variability estimates based on the high frequency data, to interpolate multiple realizations of hourly DO estimates. To conduct the GAM portion of the process as accurately to the data as possible, early testing showed that spatially limiting the data used to fit each GAM to water with similar conditions was beneficial. Therefore, our approach aimed to create segment interpolation regions that support robust DO interpolations with the right amount of data to effectively fit the GAM in target segment assessments.

The idea that spatial interpolation of DO in the Chesapeake Bay would be done on smaller regions than the entire bay is not new. The current 3D Inverse Distance Weighting-based interpolator conducts its calculations in individual segments, using a boundary region drawn to include some nearby stations outside of the target segment for informing spatial variability of DO within each assessed segment. This is a fixed process, and most users have not seen the boundary regions on a map. For the new interpolator, an initial set of segment groupings/regions is helpful to serve as an optimal performance default framework as we evaluate the other parts of the 4D interpolator along with the GAM results.

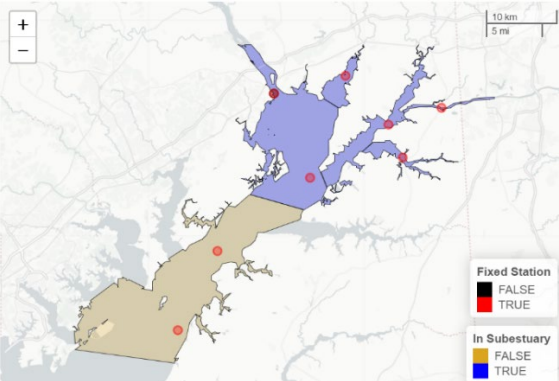
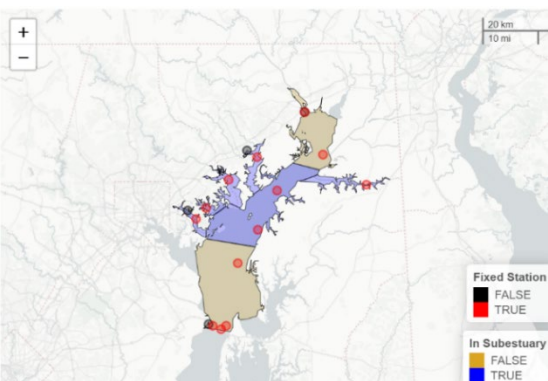
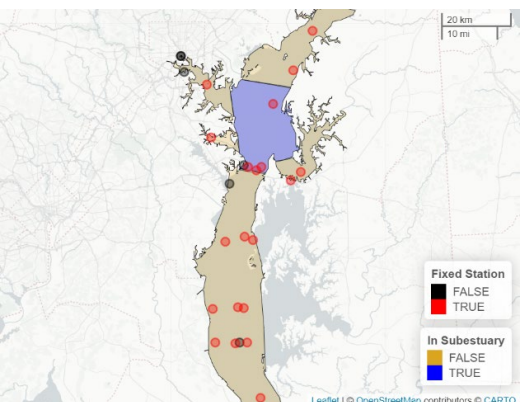
As the monitoring program evolves over time, changing and re-evaluating the performance of the default segment interpolation regions will be possible by editing an input spreadsheet, which would be an improvement over the 3D interpolator. Evaluating different combinations will also be relatively easy for an analyst by looking at diagnostic output that can be generated for different segment-region combinations. This document summarizes an initial set of segment interpolation regions and shows evaluation results of how they were selected. Although some regions cross over state lines, keep in mind that is necessary because the water is connected, and will not be too challenging with all of the preliminary work already conducted and planned by our 4D team to streamline data input each year.

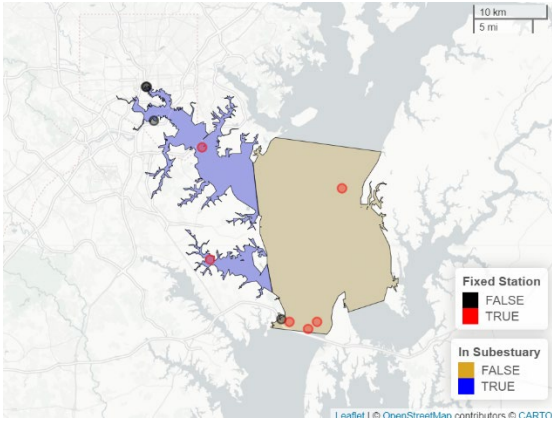
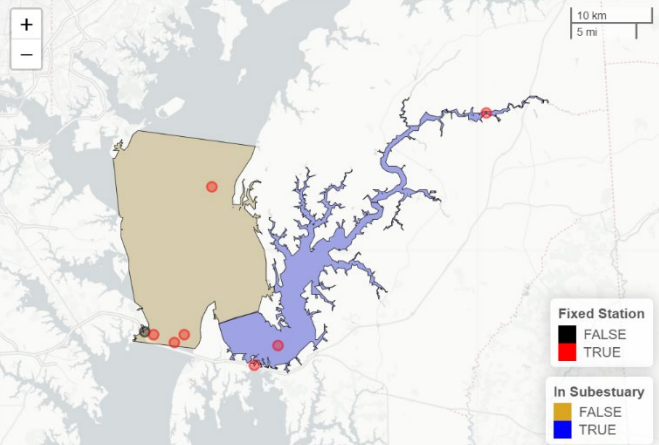
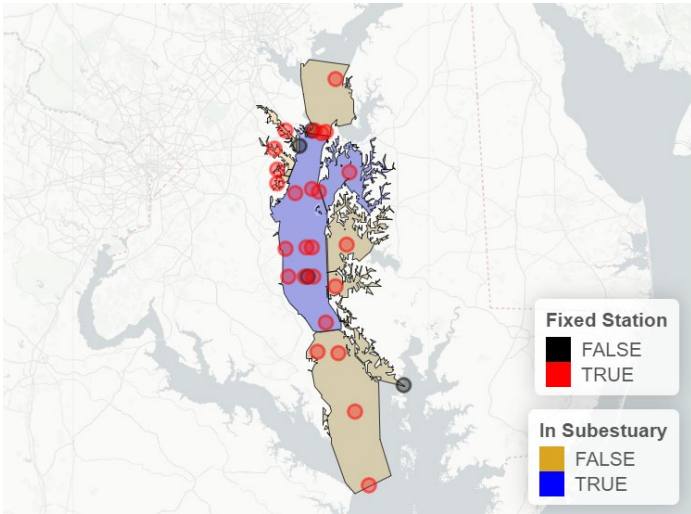
Current Segment Interpolation Regions for DO

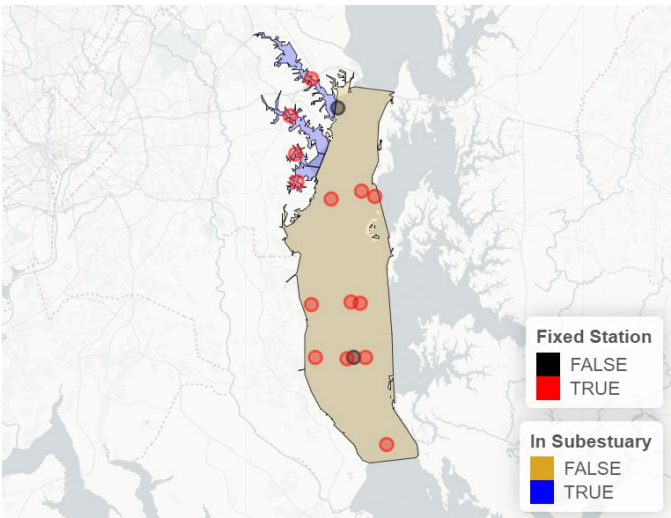
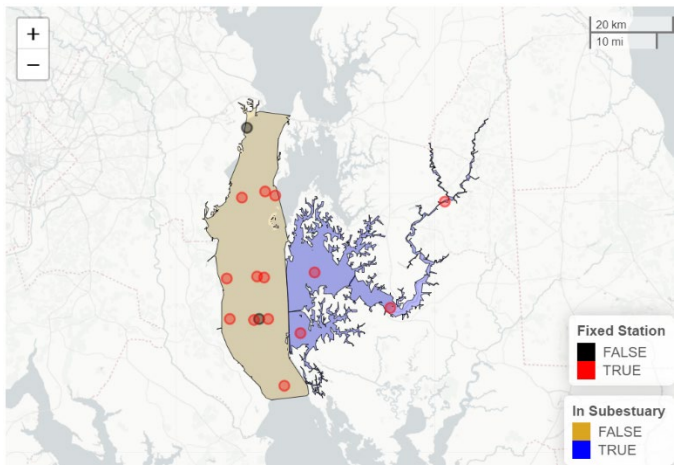
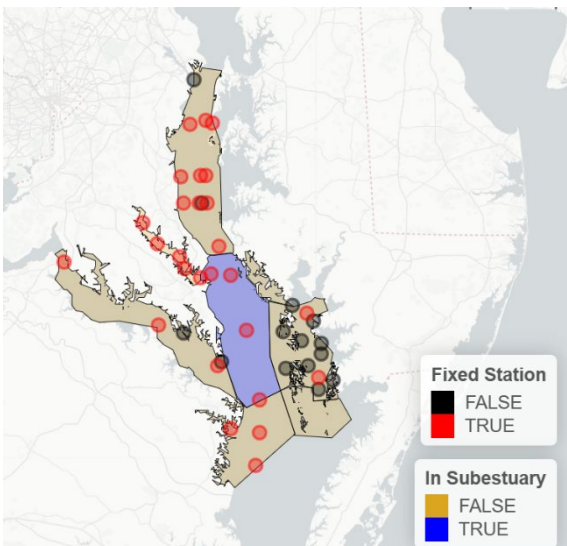
Tests were conducted to select groups of tidal segments to include together in GAM fitting. These groups of segments are called “segment interpolation regions” and the current set of these are shown in Figure 1. As of July 8, 2025, this is a draft list, and some of the regions may be adjusted based on additional testing and feedback from users. For each of these proposed regions shown in Figure 1, the purple segments are the focal segments of the group, where GAM output would be generated. The tan segments are the boundary segments included in that particular GAM fit, but not in the estimation

region. All tan segments would be primary (or purple) segments in another group. The current 31 segment interpolation regions are presented in Figure 1 with monitoring station locations from 2021, just as an example. These stations include both fixed station and shallow water continuous monitoring (common) in 2021, which might not necessarily be the same station distribution in other years. There will also be additional data used in the final 4D interpolator, including dataflow and participatory science data where it exists. Analysis of these segment interpolation regions compared to other possible segment grouping options is presented below with Figure 2 through 5.

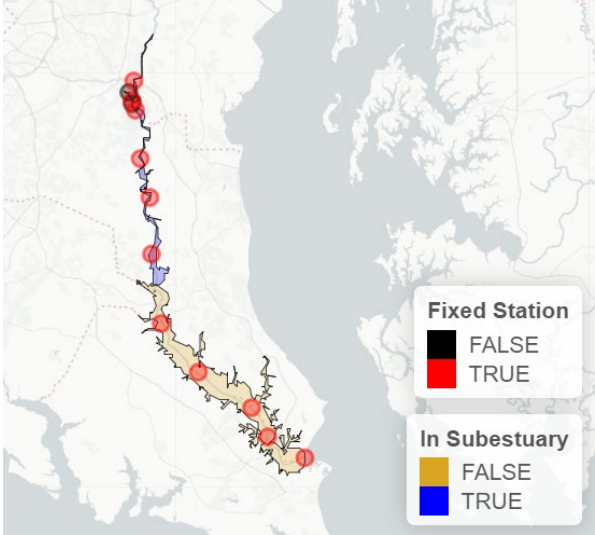
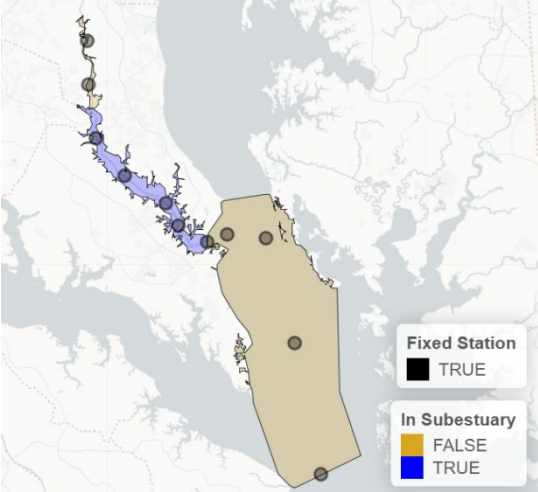
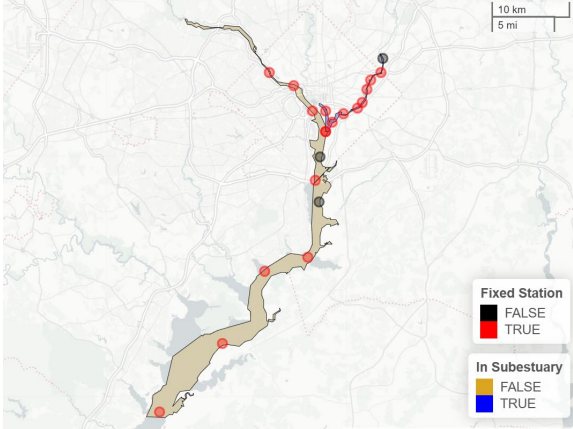
Figure 1. Segment groupings for GAM fits (version as of 7-8-2025). Purple segments are the primary focus of that region, tan segments are boundary to include all touching segments in a GAM fit. Stations shown are just an example, based on 2021 fixed station and common coverage.

Segment Interpolation Region map (2021 data)	Primary segments	Boundary
CB1_upE 	BOHOH C&DOH_DE C&DOH_MD CB1TF ELKOH NORTF	CB2OH
CB2_tribs 	BACOH BSHOH CB2OH GUNOH MIDOH SASOH	CB1TF CB3MH
CB3 	CB3MH	CB2OH CB4MH MAGMH PATMH CHSMH

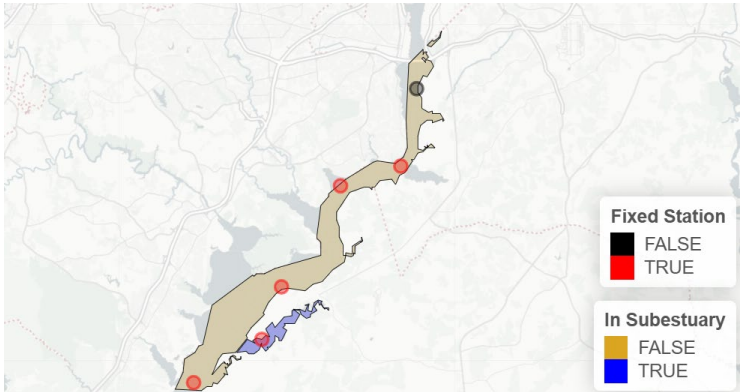

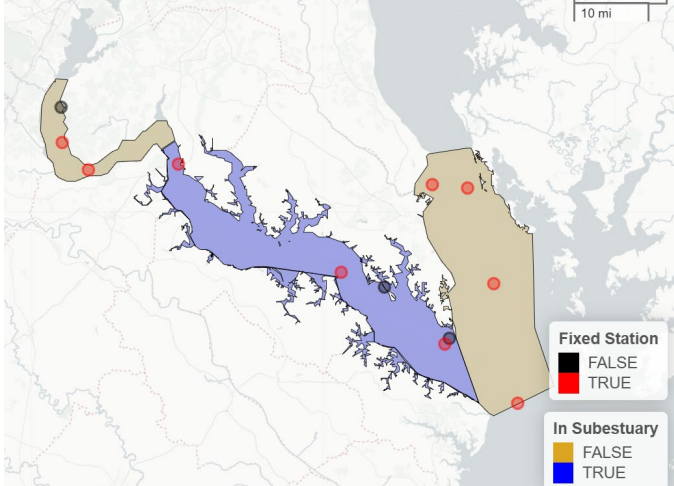
Segment Interpolation Region map (2021 data)	Primary segments	Boundary
PAT_MAG 	PATMH MAGMH	CB3MH
CHS 	CHSMH CHSOH CHSTF	CB3MH
CB4_EAS 	CB4MH EASMH	CB3MH CB5MH_MD CHOMH1 HNGMH LCHMH RHDMH SEVMH SOUMH WSTMH

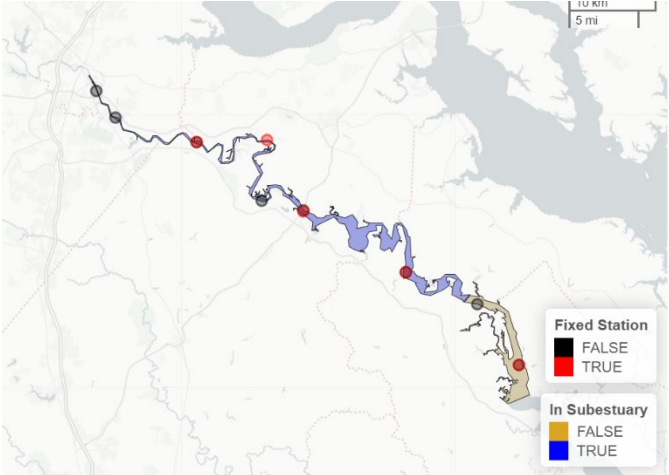
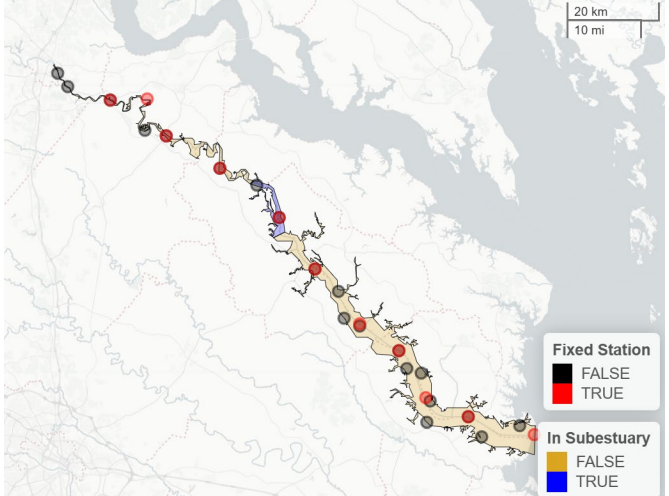
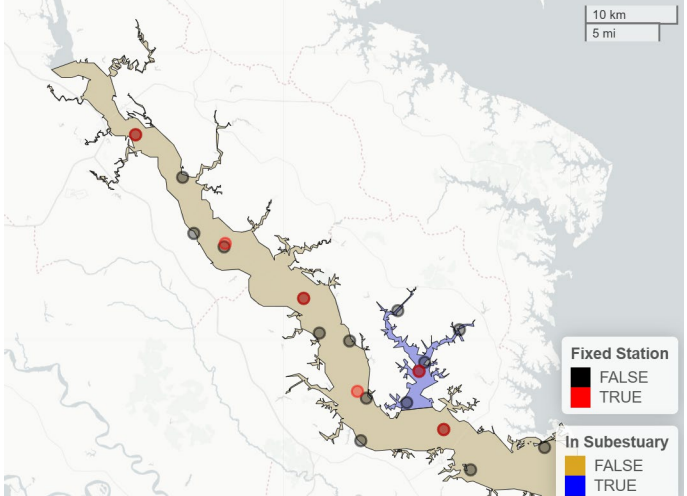
midW_MD 	WSTMH RHDMH SEVMH SOU MH	CB4MH
CHO 	CHOMH1 CHOMH2 CHOOH CHOTF LCHMH	CB4MH
CB5MD 	CB5MH_MD	CB4MH CB5MH_VA PAXMH POTMH_MD TANMH_VA TANMH_MD HNGMH *

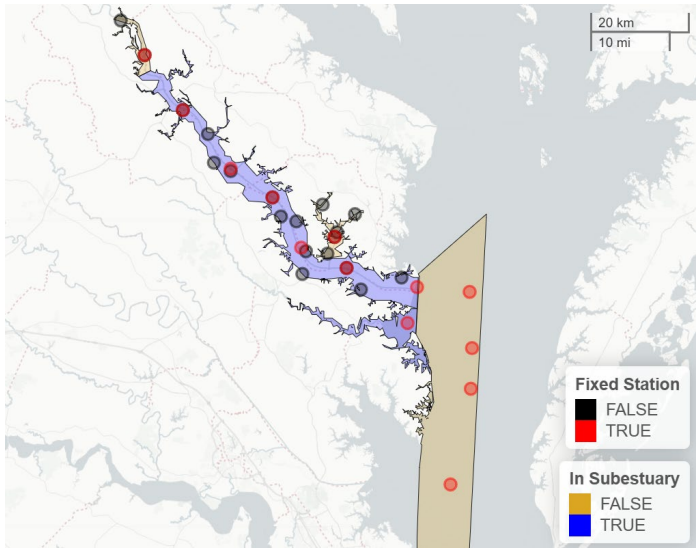
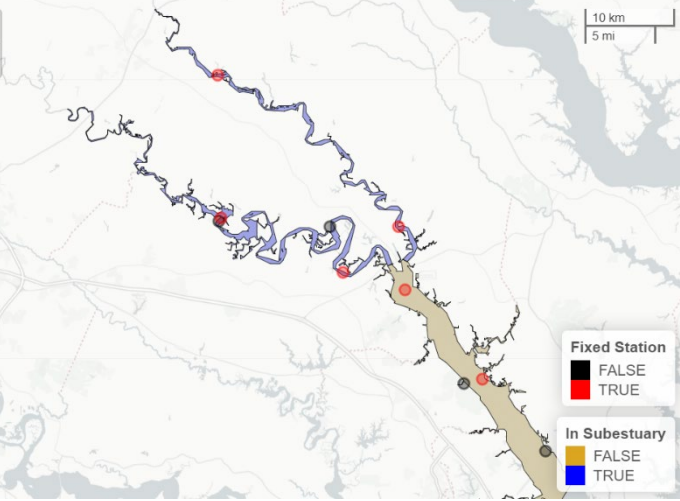
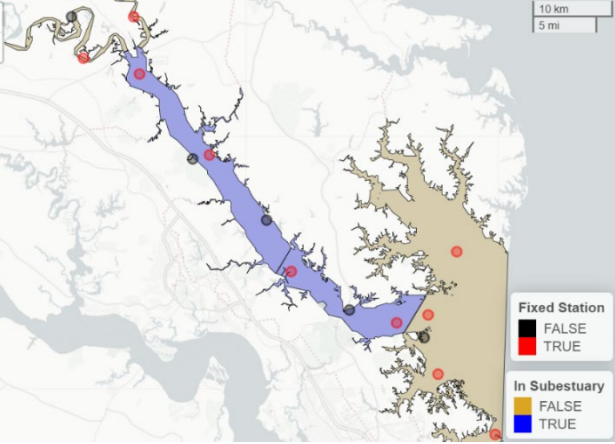
<p>CB5VA6</p> <p>Fixed Station FALSE TRUE</p> <p>In Subestuary FALSE TRUE</p>	<p>CB5MH_VA CB6PH</p> <p><i>Testing note: combining CB7PH with this group doesn't improve CB6, and increases RMSE on CB7 stations.</i></p>	<p>CB5MH_MD RPPMH PIAMH MOBPH CB8PH CB7PH TANMH_VA TANMH_MD</p>
<p>CB78</p> <p>Fixed Station FALSE TRUE</p> <p>In Subestuary FALSE TRUE</p>	<p>CB7PH CB8PH LYNPH</p>	<p>POCMH_VA TANMH_VA CB5MH_VA CB6PH JMSPH</p>

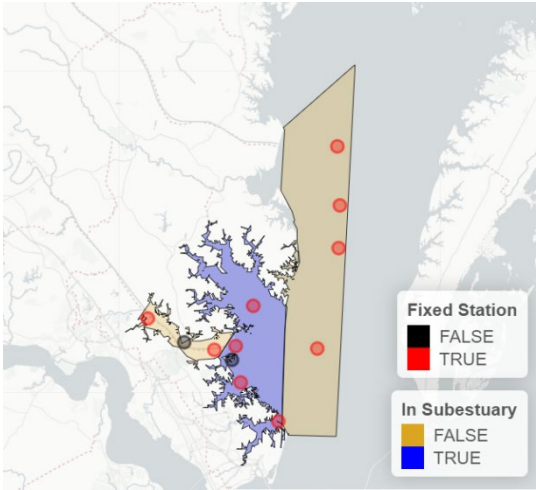
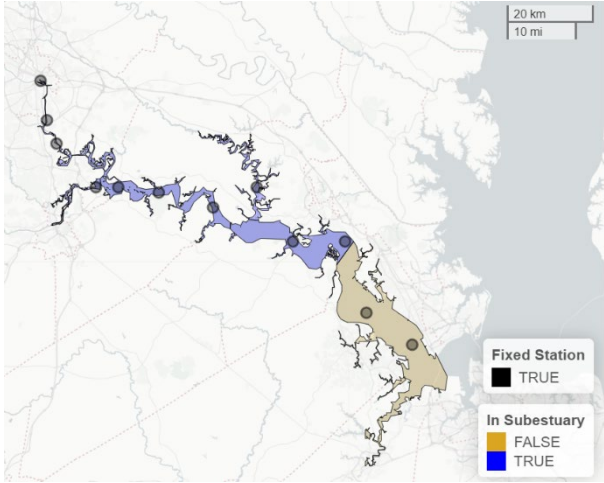
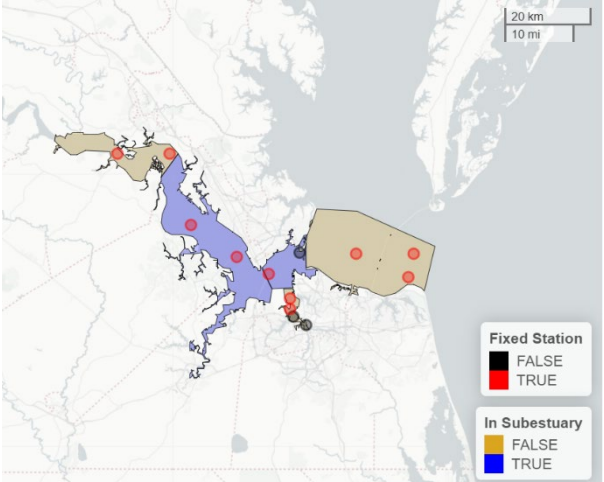
<p>upPAX</p> 	<p>PAXOH PAXTF WBRTF</p>	<p>PAXMH</p>
<p>lowPAX</p> 	<p>PAXMH</p>	<p>CB5MH_MD PAXOH</p>
<p>ANA</p> 	<p>ANATF_DC ANATF_MD</p>	<p>POTTF_DC POTTF_MD*</p> <p><i>*note: this segment is not adjacent but added to help runs happen in years with less data. Still better than joining with POTTF.</i></p>

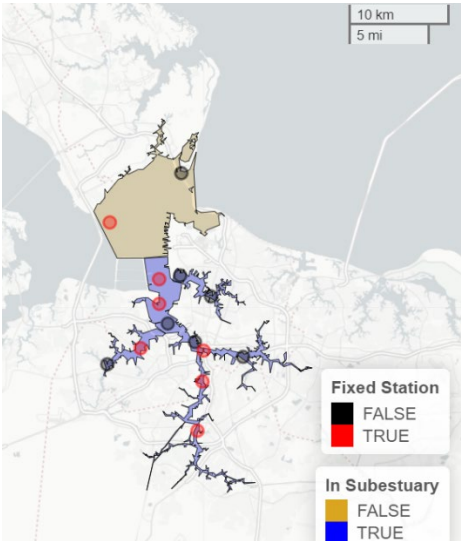
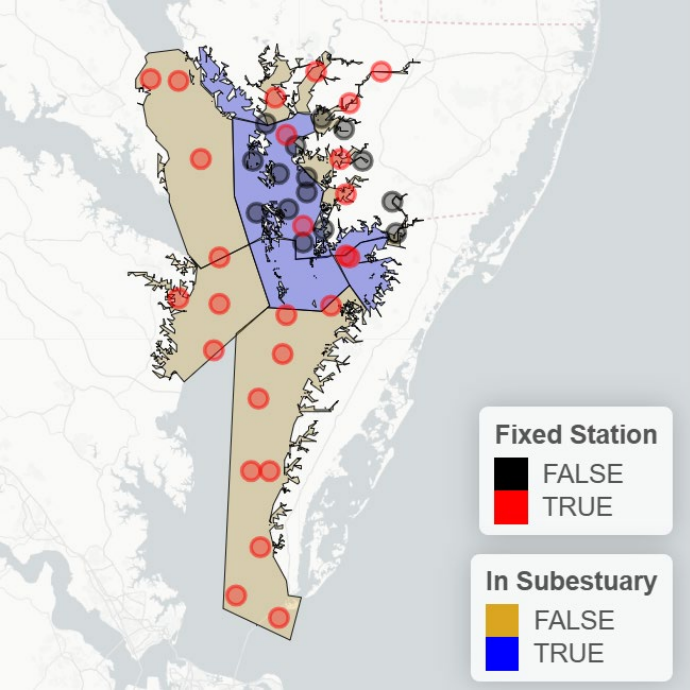
<p>POTTF</p> <p>Fixed Station FALSE TRUE</p> <p>In Subestuary FALSE TRUE</p>	<p>POTTF_MD POTTF_VA POTTF_DC</p>	<p>ANATF_DC POTOH1_MD MATTF PISTF</p>
<p>POTOH</p> <p>Fixed Station FALSE TRUE</p> <p>In Subestuary FALSE TRUE</p>	<p>POTOH1_MD POTOH_VA POTOH2_MD POTOH3_MD</p>	<p>POTMH_MD POTTF_MD</p>

<p>MATTF</p> 	<p>MATTF</p>	<p>POTTF_MD</p>
<p>PISTF</p> 	<p>PISTF</p>	<p>POTTF_MD</p>
<p>POTMH</p> 	<p>POTMH_MD POTMH_VA</p>	<p>POTOH1_MD CB5MH_MD</p>

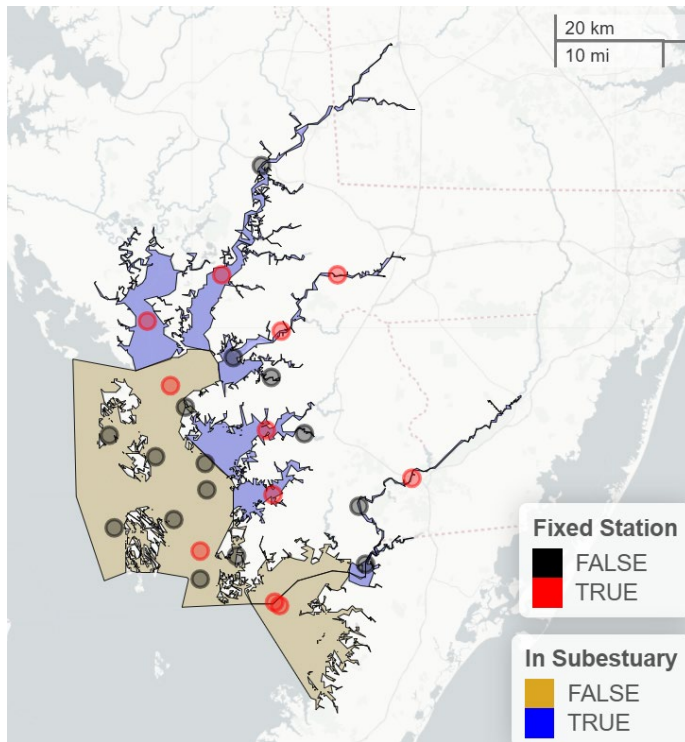
<p>RPPTF</p> 	<p>RPPTF</p>	<p>RPPOH</p>
<p>RPPOH</p> 	<p>RPPOH</p>	<p>RPPTF RPPMH</p>
<p>CRRMH</p> 	<p>CRRMH</p>	<p>RPPMH</p>

<p>RPPMH_PIAMH</p> 	<p>PIAMH RPPMH</p>	<p>RPPMH CRRMH CB6PH</p>
<p>MAT_PAM</p> 	<p>MPNOH MPNTF PMKOH PMKTF</p>	<p>YRKMH</p>
<p>lowYRK</p> 	<p>YRKMH YRKPH</p>	<p>PMKOH MPNOH MOBPH</p>

<p>MOB</p> 	<p>MOBPH</p>	<p>CB6PH YRKPH</p>
<p>upJMS</p> 	<p>JMSTF1 JMSTF2 APPTF CHKOH JMSOH</p>	<p>JMSMH</p>
<p>lowJMS</p> 	<p>JMSPH JMSMH</p>	<p>CB8PH ELIPH JMSOH</p>

<p>ELZ</p> 	<p>ELIPH LAFMH EBEMH SBEMH WBEMH</p>	<p>JMSPH</p>
<p>TAN_POCMH</p> 	<p>TANMH_MD TANMH_VA HNGMH POCMH_MD POCMH_VA</p>	<p>CB5MH_MD CB5MH_VA CB7PH POCOH_VA POCOH_MD BIGMH MANMH WICMH NANMH FSBMH</p>

LowE



FSBMH
NANMH
NANOH
NANTF_MD
NANTF_DE
WICMH
MANMH
BIGMH
POCOH_MD
POCOH_VA
POCTF

TANMH_MD
POCMH_MD
POCMH_VA

Evaluation

Several iterations of possible segment interpolation regions were conducted in order to settle on the 31 groups shown in Figure 1. Figure 2 shows three example sets of groups tested.

- Figure 2a) Whole bay together, with no segment delineations. The tidal waters fit together in one GAM.
- Figure 2b) Relatively large segment groups. Note that for these 11 groups, all adjacent boundary segments were used as additional data to fit each of the 11 GAMs (similar to the structure shown in Figure 1).
- Figure 2c) Small, or no, segment groups. It would be reasonable to consider fitting a unique GAM to each of the 92 tidal segments, including boundary data. This test was done as shown in Figure 2c. Note that 79 groups were tested instead of 92 because it was already clear that very small station groups would not be sufficient to fit a GAM and would fail due to lack of data. The GAM requires enough data to fully define the relationship between DO and each of the explanatory variables. So in some cases, small segments that are split only due to them spanning a state line (e.g., ANATF, POTTF, NANTF, POCOH) were combined. However, even with these logical combinations, many of these individual segments failed for the GAM fit due to lack of data. Because common data is not always in the same place every year, a test was done with only the fixed stations, and the regions that failed with only that data are indicated with black shading in Figure 2c. For these places at a minimum, more grouping would be needed. Those considerations led to slightly larger groups in Figure 2d.
- Figure 2d) Current selection of 31 segment groups. Additional tests were done besides the three comparisons shown to settle on these exact groups.

Note that the GAM model used for these tests were fit to daily data. Work is underway to use all hourly data in GAM fitting, but it is not expected that the change will impact the segment interpolation region selection.

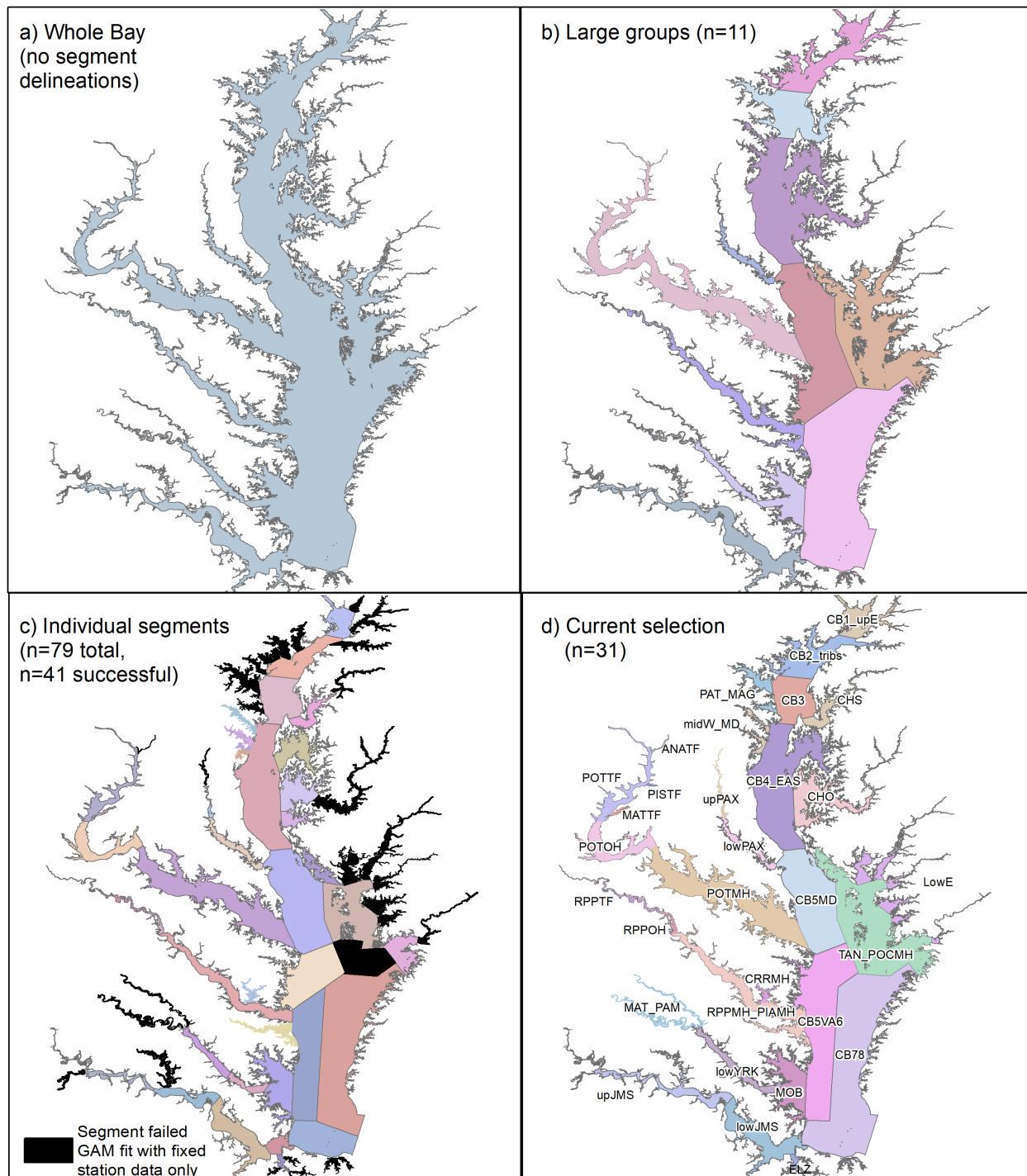


Figure 2. Different segment interpolation region options tested for fitting GAM portion of 4D interpolation. Note that these colored regions are only the primary segments grouped, but for panels b, c, and d, every colored region also includes the touching boundary segments as boundary data. Black segments in 2c failed GAM fits due to insufficient data. Panel d matches Figure 1 and are referred to as the selected set or groups below.

Each segment group shown in Figure 2 was used to fit GAMs to the 2021 DO data. Table 2 shows overall R^2 for the 2021 GAM fits. The whole bay GAM has lower R^2 than most of the grouped GAMs. Figure 3 shows the root mean squared error (RMSE) computed at each station ($n=236$) from options a-c on the x-axis versus the currently selected groups (Figures 1 and 2d) on the y-axis. These results show general improvement both with R^2 and RMSE when less grouping is used. Figure 4 shows a comparison between the selected set of groups (Figure 1) and the other options spatially by station.

Table 2. GAM R^2 range for each option

Grouping	R^2 2021
a) Whole bay: one GAM	0.74
b) More grouped: 11 GAMs	0.77-0.93
c) Individual segments: 79 possible	0.69-0.99
d) Selected: 31 GAMs	0.69-0.98

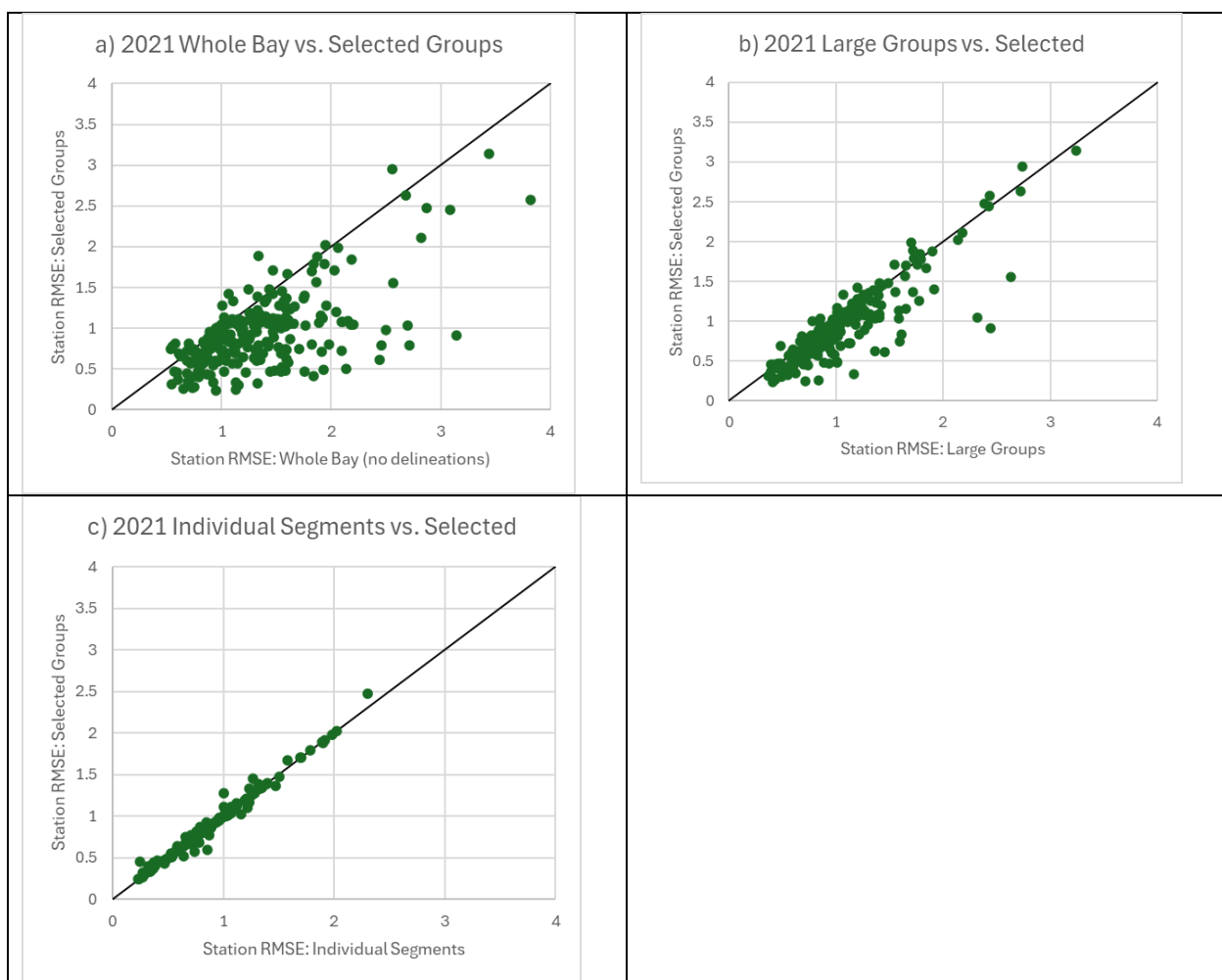


Figure 3. RMSE of the GAM estimates at each station for 2021 data comparing various segment groups to fit the GAM. The x-axis in each case is a different option plotted vs. the y-axis results that are the currently selected grouping show in Figure 1 and Figure 2d ($n=31$ groups). If a circle is below the black line, it indicates the selected set of regions (Figure 1) results in less error at that station. Station counts are: $n=236$ for a-b and $n=163$ for c.

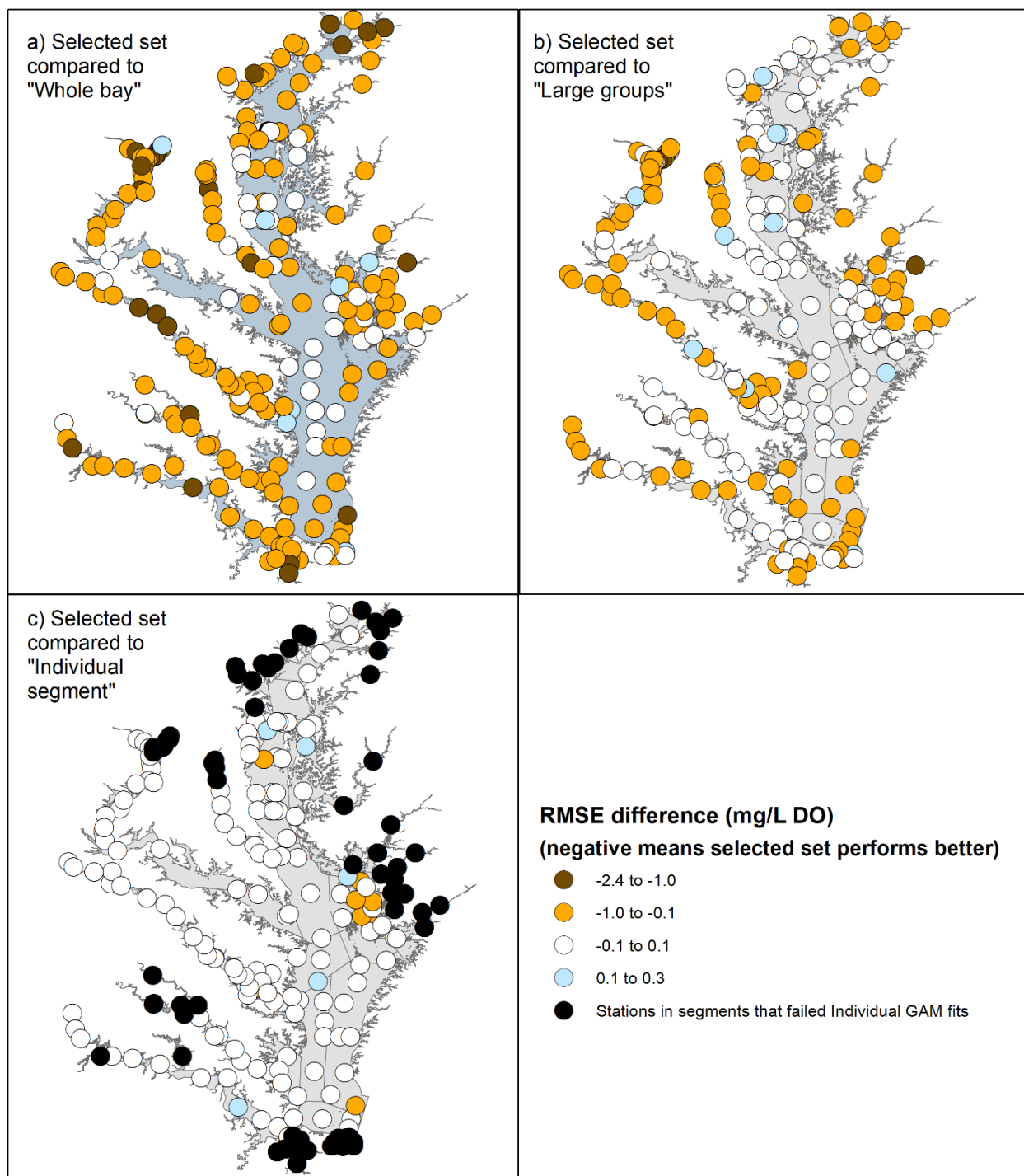


Figure 4. Difference in RMSE of 2021 GAM fits compared to the data mapped by station. Each panel shows the difference between the RMSE using the selected segment grouping (Figure 1 and Figure 2d) compared to another option (a-c).

Analysis of Figure 4 supports the general conclusion that less grouping is better for accurate GAM estimates. There is a feasibility limit though, with regions needing to have a minimum amount of data to fit the GAM. Thus, a balance in the selected regions was made between the “Large Groups” (b) and “Individual Segments” (c). Note, there are several regions where we made the choice with the selected set to be fairly small (Rappahannock and Potomac Rivers). However, based on other work, we suspect that a possible problem with small regions may arise when it is necessary to generate GAM estimates on the interpolation grid at points beyond the spatial bounds of the data. We will take a close look at results from every part of the development process with this in mind, and may make further changes to these regions.

For a final evaluation, it is possible to zoom in on the GAM estimates alongside the observed data at the stations to further understand the impacts of different levels of spatial grouping on GAM estimates. ET10.1 in the Pocomoke tidal fresh river is the furthest station on the east side of the maps (southern eastern shore tributary). When the entire bay is included in one group (Figure 2a), the GAM estimates at this station, shown with the black lines, over predict the DO observed at 1m depth greatly (Figure 5a). When the POCTF segment is grouped slightly more in the 11 Large Groups (Figure 5b), the GAM gets closer to fitting the data. But not until the lower eastern shore tributaries are grouped together as the current LowE group with only the Tangier Sound segments as boundary do the GAM estimates get most accurate to the middle of the data at the two depths shown (Figure 5c). This is an example that is similar to many parts of the bay where it is better to manage the regions so that large mainstem segments are not included as boundary segments to smaller tributaries.

Figure 5a. “Whole bay” ET10.1, 2021 using one bay-wide GAM (RMSE=1.4 at this station). Red dots are observations, black lines are GAM estimates at each depth.

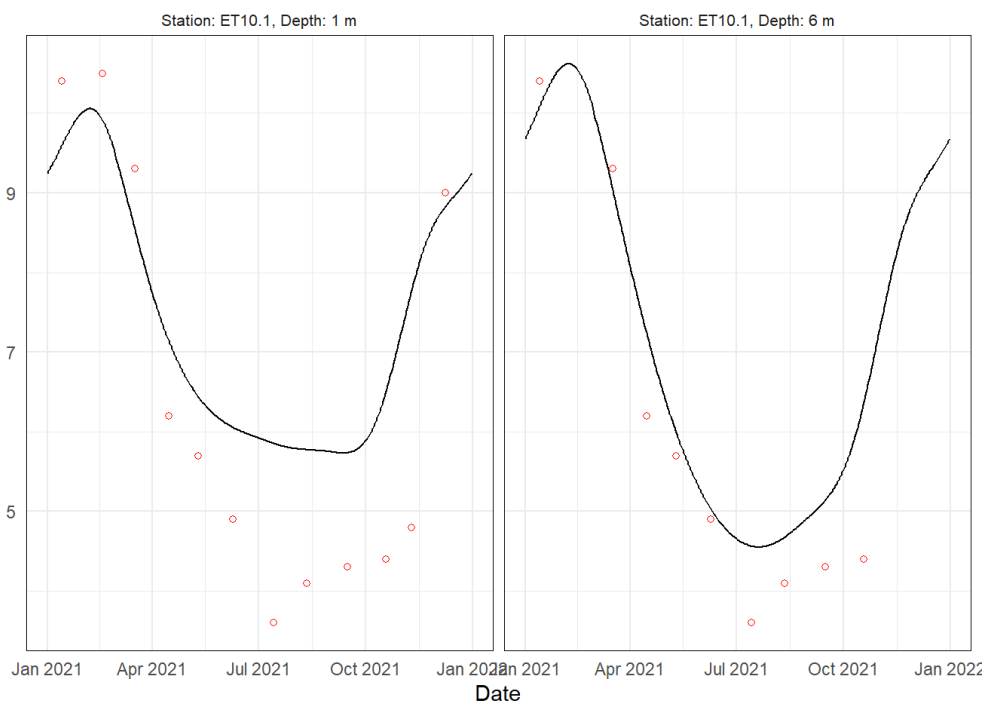


Figure 2b. “Large groups” ET10.1, 2021, using more grouped option that combines Tangier and Eastern Shore tribs + boundaries of CB7PH and CB5MH_VA (RMSE=0.80)

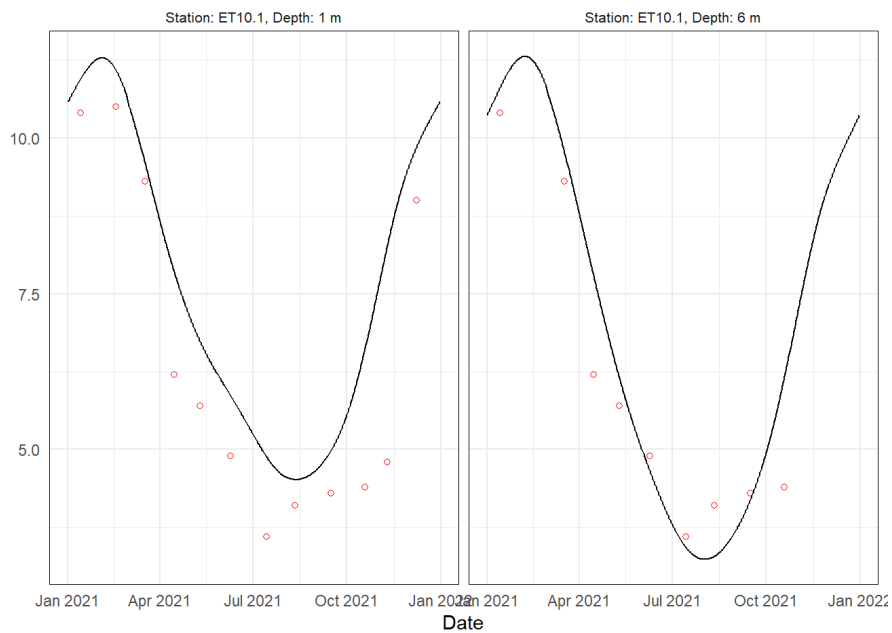
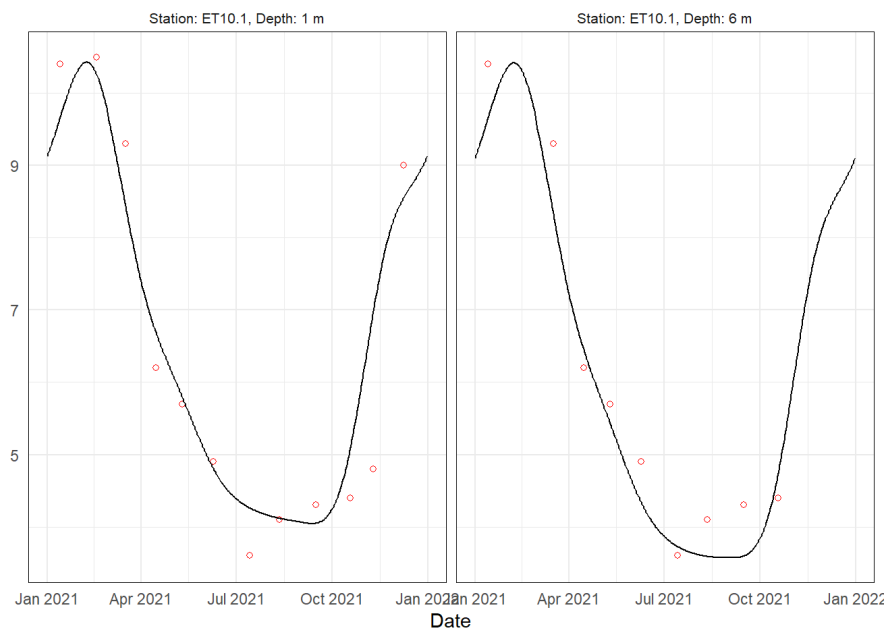


Figure 2c. “Selected set” ET10.1, 2021, that is only Eastern shore tribs and boundary as Tangier segments, LowE (RMSE=0.69). *Note: this station does not have a result for “individual segment” option.*



Next steps

Evaluation of these regions will continue as we generate GAM daily estimates on the interpolation grid using all high frequency data and combine the hourly estimates. Feedback is welcome on these groups as well as what information would be helpful to understand them.