Appendix #. Stability of Dissolved Oxygen Concentration Threshold Estimates Across Simulation Counts
Updated: August 2, 2026 **Draft**

#-1. Summary

This analysis evaluates the agreement between estimates derived from fewer than 100 dissolved oxygen (DO) simulations and the corresponding estimates derived from the full 100-simulation set. Each simulation provides one realization of the segment-level percentage of space-time grid-hours below the applicable DO threshold. The analysis repeatedly draws random subsets of different sizes from the 100 simulations and compares the subset P10[1], mean, and P90 with the corresponding full-set statistics. It evaluates both displacement from the 100-simulation reference and the P95–P05 range across repeated subset draws.

Segments were assigned to mutually exclusive analysis classes to prevent the same segment from influencing more than one class summary. The analysis therefore includes 10 Deep Channel (DC) segments, 8 additional Deep Water (DW) segments, and 74 remaining Open Water (OW) segments. These counts describe the non-overlapping analysis classes, not the total number of segments to which each designated use applies.

Increasing the number of simulations generally reduced displacement from the 100-simulation reference and narrowed the repeated-draw P95–P05 range. The mean generally stabilized most rapidly, whereas P10 and P90 required more simulations to achieve comparable reproducibility. Improvements continued as the ensemble increased but generally became progressively smaller around 50–60 simulations.

At 50 simulations, the class-median absolute deviation from the 100-simulation reference was less than 0.0001 percentage point for P10 in all three classes, ranged from 0.0004 to 0.0007 percentage points for the mean, and ranged from 0.0000 to 0.0001 percentage point for P90. The corresponding class-median P95–P05 repeated-draw ranges were 0.012–0.037, 0.021–0.026, and 0.035–0.044 percentage points, respectively. Across the three classes, the largest segment-level absolute deviations were 0.0766 for P10 (CHSTF), 0.1306 for the mean (ANATF_MD), and 0.3004 for P90 (ANATF_MD); the largest repeated-draw ranges were 0.576, 1.510, and 5.073 percentage points, all for ANATF_MD. These results quantify the reproducibility provided by 50 simulations; they are not presented as a formal convergence threshold.

For testing and exploratory work, 21 simulations may be adequate when the purpose is to evaluate workflow behavior rather than support a final criteria determination. For production planning, approximately 50–60 simulations provide a reasonable provisional operating range based on the observed diminishing returns in stability. A 51-simulation ensemble also has the practical property that P10, the median, and P90 correspond directly to the 6th, 26th, and 46th ordered simulation-level results.

This analysis provides a defensible empirical basis for selecting simulation counts for testing, planning, and workflow development. Future sensitivity analyses should incorporate additional spatial measures and other indicators of simulation stability or assessment performance to help determine the final number of simulations used for assessment. The production count should also be evaluated using the CFD procedure and percentile-based decision rule ultimately adopted because they represent the final assessment endpoint.

ANATF_MD remains a documented high-sensitivity stress case within the analysis. Its instability is consistent with a combination of comparatively limited grid-hour support and a large share of interpolated mean DO values near the 3.2 mg/L threshold. Under these conditions, moderate correlated simulation departures can move many grid-hours across the threshold. ANATF_MD should therefore receive a segment-specific convergence check rather than be allowed to determine the general simulation-count recommendation for otherwise more stable segments. Its behavior is examined separately in **Exhibit #-2**.

[1] Pxx denotes the xxth percentile. P10 and P90 are the 10th and 90th percentiles of the simulation-level percent-below-threshold results. P05 and P95 describe the 5th and 95th percentiles across repeated random subset draws; therefore, the P95–P05 range measures the reproducibility of a statistic estimated from a given number of simulations.

#-2. Method

The analysis was structured as a repeated-subset experiment. For each segment, the 100 baseline dissolved oxygen (DO) simulations were treated as the reference population. The evaluated DO thresholds represent management-relevant endpoints associated with the Deep Channel, Deep Water, and Open Water designated uses in the Chesapeake Bay water-quality standards attainment framework.

For each simulation, segment, and applicable DO threshold, the raw simulation results were reduced to one segment-level quantity: the percentage of simulated space-time grid-hours with DO below the threshold. Each simulation therefore contributed one percent-below-threshold result. Subsets of different sizes were repeatedly drawn from the 100-simulation population, and the P10, mean, and P90 of the simulation-level results within each subset were compared with the corresponding statistics calculated from all 100 simulations.

The experiment evaluated how displacement from the 100-simulation reference and variability among repeated subset draws changed as the number of simulations increased. The 100-simulation results provide a common empirical reference for these comparisons; they are not assumed to represent theoretical truth. Accordingly, the analysis was designed to identify diminishing improvements in stability rather than apply a predetermined convergence threshold.

#-2.1 Analysis scope

To avoid allowing the same segment to influence more than one class summary, segments were assigned to mutually exclusive analysis classes. Segments containing a Deep Channel designated use were assigned to the Deep Channel class, the remaining segments containing a Deep Water designated use were assigned to the Deep Water class, and all remaining segments were assigned to the Open Water class. The resulting counts—10 Deep Channel, 8 Deep Water, and 74 Open Water segments—therefore describe the non-overlapping analysis classes rather than the complete inventories of segments to which each designated use applies.

The class-specific DO thresholds were 1.0 mg/L for Deep Channel, 1.7 mg/L for Deep Water, and 3.2 mg/L for Open Water. Candidate subset sizes were 1, 5, 10, 20, 30, 40, 50, 60, 70, and 80 simulations. For each candidate size, 100 subsets were randomly drawn from the 100-simulation population. Simulations were sampled without replacement within each subset; however, because the repetitions were independent, an individual simulation could be included in multiple subsets. A fixed random seed made the sampling experiment reproducible.

Table #-1 identifies the complete inventory of segments assigned to each analysis class. **Table #-2** summarizes the repeated-subset design.

Table #-1. Mutually exclusive analysis classes, dissolved oxygen thresholds, and assigned segments.

Designated-use class	Dissolved oxygen threshold (mg/L)	Segment IDs / note
Deep Channel	1.0	10 segments: CB3MH, CB4MH, CB5MH_MD, CB5MH_VA, CHSMH, EASMH, PATMH, POTMH_MD, POTMH_VA, RPPMH
Deep Water	1.7	8 segments: CB6PH, CB7PH, MAGMH, PAXMH, SBEMH, SEVMH, SOUMH, YRKPH
Open Water	3.2	74 segments: ANATF_DC, ANATF_MD, APPTF, BACOH, BIGMH, BOHOH, BSHOH, C&DOH_DE, C&DOH_MD, CB1TF, CB2OH, CB8PH, CHKOH, CHOMH1, CHOMH2, CHOOH, CHOTF, CHSOH, CHSTF, CRRMH, EBEMH, ELIPH, ELKOH, FSBMH, GUNOH, HNGMH, JMSMH, JMSOH, JMSPH, JMSTF1, JMSTF2, LAFMH, LCHMH, LYNPH, MANMH, MATTF, MIDOH, MOBPH, MPNOH, MPNTF, NANMH, NANOH, NANTF_DE, NANTF_MD, NORTF, PAXOH, PAXTF, PIAMH, PISTF, PMKOH, PMKTF, POCMH_MD, POCMH_VA, POCOH_MD, POCOH_VA, POCTF, POTOH_VA, POTOH1_MD, POTOH2_MD, POTOH3_MD, POTTF_DC, POTTF_MD, POTTF_VA, RHDMH, RPPOH, RPPTF, SASOH, TANMH_MD, TANMH_VA, WBEMH, WBRTF, WICMH, WSTMH, YRKMH

Table #-2. Monte Carlo subset-draw design.

Analysis element	Specification
Reference simulation set	100 simulations per segment-threshold combination
Reference statistics	P10, Mean, and P90 computed from the full 100-simulation set
Candidate subset sizes	1, 5, 10, 20, 30, 40, 50, 60, 70, and 80 simulations
Number of repeated subset draws per candidate size	100
Sampling scheme	Random sampling without replacement from the 100-simulation reference set
Statistics recalculated for each subset draw	P10, Mean, and P90
Central summary across repeated subset draws	Median
Repeated-draw percentile interval	5th to 95th percentile
Derived deviation metric	Subset-draw median minus full-simulation reference value
Derived percentile range metric	P95 minus P05 across repeated subset draws
Random seed	Fixed seed used for reproducibility
Unit of analysis	Segment-threshold combination

#-2.2 Decision-relevance screening

Segments with negligible below-threshold responses provide little information about the stability of P10, the mean, or P90. Such segments were retained in the complete inventory but excluded from the class figures and summary tables used to evaluate simulation-count stability.

All 10 Deep Channel segments and all 8 Deep Water segments were retained. Among the 74 Open Water segments, APPTF had no grid-hours below 3.2 mg/L in any of the 100 simulations. NORTF had a below-threshold response in only one simulation, and its maximum percentage below the threshold was approximately 0.0001%. These two segments were therefore excluded from the Open Water stability summaries but remain listed in **Table #-1.**

Consequently, the Deep Channel, Deep Water, and Open Water class summaries represent 10, 8, and 72 segments, respectively. The Open Water figure and the Open Water entries in **Table #-7** and **Table #-8** are based on the 72 retained Open Water segments.

#-2.3 Full-simulation reference set

For each simulation, segment, and threshold, the raw simulation output was reduced to one segment-level quantity: the percentage of simulated space-time grid-hours falling below the threshold. The 100 simulations therefore produced a 100-element vector of percent-below-threshold values for each segment-threshold combination. The P10, mean, and P90 of this vector were used as the reference statistics.

#-2.4 Repeated subset-draw experiment

For each candidate subset size, n, the analysis selected n distinct simulation-level values without replacement from the 100-simulation reference set. The P10, mean, and P90 were then recalculated from the selected values. This process was repeated 100 times for each subset size. Each repetition began with the full set of 100 simulations; consequently, although a simulation could be selected only once within a given subset, it could be selected again in other repetitions. The experiment therefore produced 100 subset estimates for each statistic and subset size.

Each repeated-draw distribution was summarized by its median, 5th percentile (P05), and 95th percentile (P95). Here, the median represents the central result across the 100 repeated subset draws. The following two measures were then calculated:

Deviation from reference = median subset estimate − 100-simulation reference estimate

Repeated-draw range = P95 subset estimate − P05 subset estimate

Deviation from reference measures the centering of the repeated subset estimates relative to the 100-simulation result. A positive value indicates that the median subset estimate was above the reference, whereas a negative value indicates that it was below the reference. The repeated-draw range measures how much the estimated statistic varied among different subsets containing the same number of simulations.

The class figures retain signed deviation so that the direction of the deviation remains visible. **Table #-7** summarizes the class median of the absolute segment-level deviations, thereby preventing positive and negative deviations from canceling one another. **Table #-8** summarizes the class median of the segment-level P95–P05 repeated-draw ranges.

#-2.5 Transition from segment-level results to class-level interpretation

Figure #-1 shows how the repeated-subset calculations were translated from an individual-segment result into the class-level displays used in this analysis. The figure is organized into three columns corresponding to P10, the mean, and P90 of the simulation-level percent-below values. The top portion presents results for one illustrative segment, while the bottom portion presents the corresponding stability measures for all segments within the designated-use class.

In the top portion, each point represents the median estimate obtained from the 100 repeated subset draws at a given simulation count. The dashed horizontal line shows the corresponding statistic calculated from all 100 simulations, and the vertical interval extends from P05 to P95 across the repeated draws. The distance between the point and the dashed line indicates relative accuracy, or how closely the median subset estimate is centered on the 100-simulation reference. The length of the P05–P95 interval indicates precision, or how consistently different subsets of the same size reproduce the result.[2]

P10 and P90 should not be confused with P05 and P95. P10 and P90 are statistics calculated from the simulation-level percent-below values within each subset. P05 and P95 describe the distribution of estimates obtained after repeating the subset selection 100 times. The resulting P95–P05 range therefore measures sensitivity to which simulations were selected.

The bottom portion carries these same measures into the class-level display. The columns again show P10, the mean, and P90. The upper row plots each segment's signed deviation from its 100-simulation reference and therefore summarizes relative accuracy. The lower row plots each segment's P95–P05 repeated-draw range and therefore summarizes precision. The red highlighted line identifies the segment with the largest repeated-draw range at $n = 10$ within the analysis class for the statistic shown, as named in the panel label. This labeling facilitated follow-up investigations as needed.

As the simulation count increases, deviations are expected to approach zero and repeated-draw ranges are expected to narrow. A result can be closely centered on the reference while still varying substantially among repeated subsets; conversely, repeated subsets can produce a narrow range that remains displaced from the reference. Both measures therefore contribute to evaluating an adequate simulation count. Because P10 and P90 describe the tails of the simulation-level distribution, they generally require more simulations than the mean to achieve comparable precision. Because the number of violations allowed in a segment is typically small, precision in the tails of the distribution is more relevant than precision in the center of the distribution.

[2] **Exhibit #-1** presents the complete illustrative analysis underlying **Figure #-1** and this interpretation.

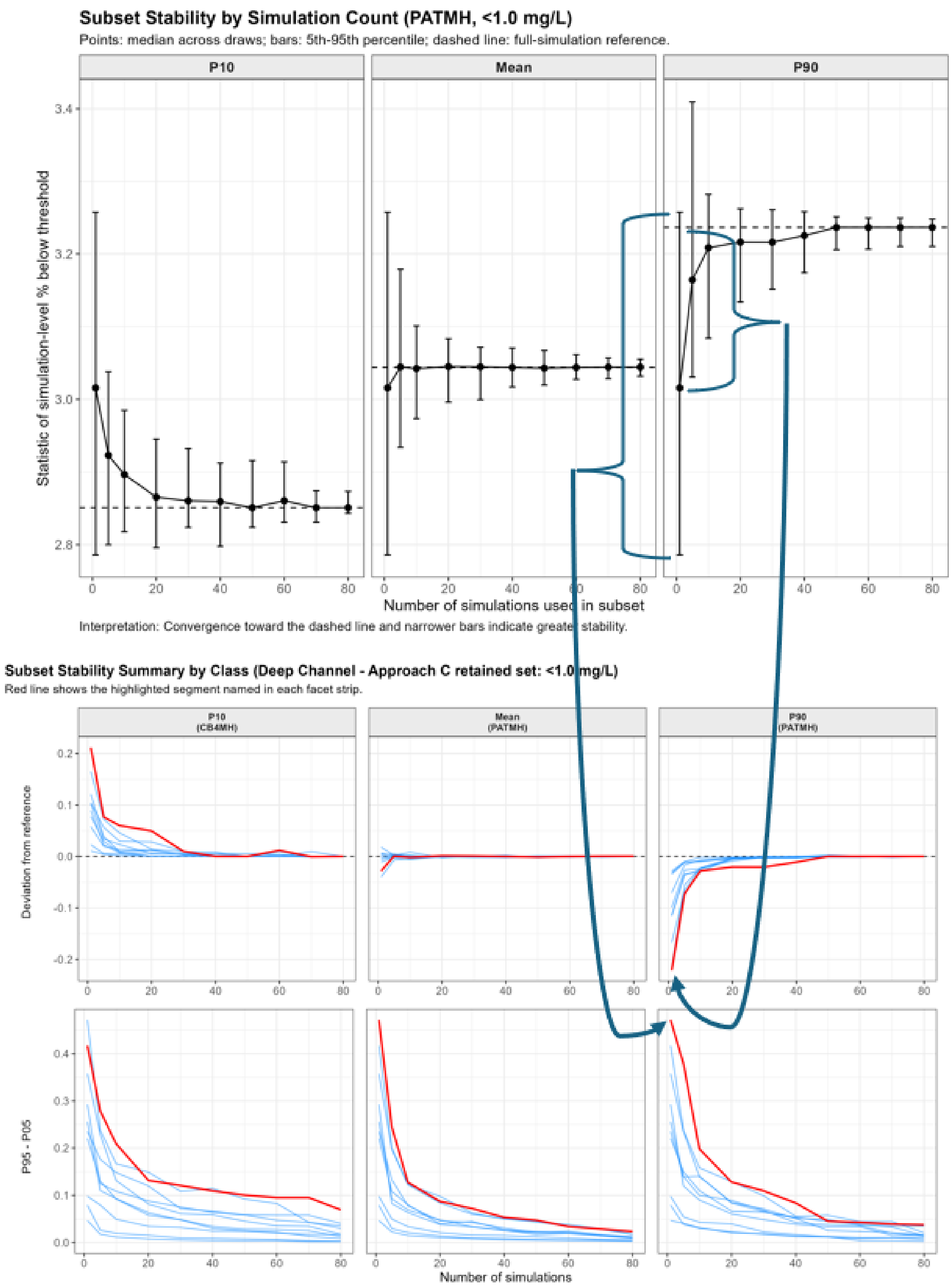

Figure #-1. Relation between segment-level subset-stability results and class-level ensemble line summaries. Top: illustrative segment-level subset-stability plot. Bottom: corresponding class-level ensemble plot. The top row of the class display shows deviation from the segment reference; the bottom row shows the P95–P05 repeated-draw range.

#-3. Results

The class results are presented first so the segment trajectories can be examined before the cross-class comparison. In each class section, the figure shows all retained segment trajectories and the companion table reports both the class median and maximum at selected simulation counts. The median represents the typical segment, while the maximum shows the most sensitive segment at that statistic and simulation count.

#-3.1 Deep Channel results

Deep Channel showed the most orderly class-level behavior. **Figure #-2** shows the trajectories for all 10 retained segments. The repeated-draw ranges narrowed steadily as the number of simulations increased, and no segment remained persistently separated from the class pattern across all three statistics.

At 20 simulations, the median P95–P05 ranges were 0.072 percentage points for P10, 0.043 for the mean, and 0.075 for P90; the corresponding maxima were 0.149 for PATMH, 0.098 for CB4MH, and 0.129 for CB4MH (**Table #-3**). At 50 simulations, the medians had declined to 0.037, 0.021, and 0.035, while the maxima were 0.100 for CB4MH, 0.048 for PATMH, and 0.058 for CB4MH. By 60 simulations, the medians were 0.031, 0.020, and 0.035, and the maxima were 0.095, 0.038, and 0.045. The comparatively modest change from 50 to 60 simulations is consistent with diminishing improvements in stability for the Deep Channel class.

#-3.2 Deep Water results

Deep Water had similarly small median deviations but a wider spread among segments, particularly for P90. **Figure #-3** shows that PAXMH, highlighted in red, accounts for much of the upper end of the repeated-draw ranges.

At 20 simulations, the median P95–P05 ranges were 0.056 percentage points for P10, 0.060 for the mean, and 0.088 for P90; the corresponding maxima were 0.307, 0.224, and 0.395, all for PAXMH (**Table #-4**). At 50 simulations, the medians had declined to 0.025, 0.026, and 0.037, while the maxima were 0.121, 0.103, and 0.221, again all for PAXMH. By 60 simulations, the medians were 0.020, 0.024, and 0.032; the maxima were 0.093 for MAGMH, 0.077 for SEVMH, and 0.217 for PAXMH. The P90 maximum therefore changed little between 50 and 60 simulations and remained the principal Deep Water tail-sensitivity result.

#-3.3 Open Water results

The 72-segment Open Water retained set had small class medians but the greatest segment-level heterogeneity. **Figure #-4** shows the typical narrowing of the class trajectories together with the much wider ANATF_MD trajectory highlighted in red.

At 20 simulations, the median P95–P05 ranges were 0.036 percentage points for P10, 0.040 for the mean, and 0.080 for P90; the corresponding maxima were 1.061 for CHSOH, 3.564 for ANATF_MD, and 10.154 for ANATF_MD (**Table #-5**). At 50 simulations, the medians had declined to 0.012, 0.021, and 0.044, while the maxima were 0.576, 1.510, and 5.073, all for ANATF_MD. By 60 simulations, the medians were 0.011, 0.017, and 0.045, and the ANATF_MD maxima were 0.548, 1.628, and 4.837. Small nonmonotonic changes between adjacent simulation counts are possible because P05 and P95 were estimated from 100 repeated subset draws; they do not alter the overall narrowing pattern. The class median describes the typical Open Water segment, whereas the much larger ANATF_MD mean and P90 ranges identify a distinct high-sensitivity case.

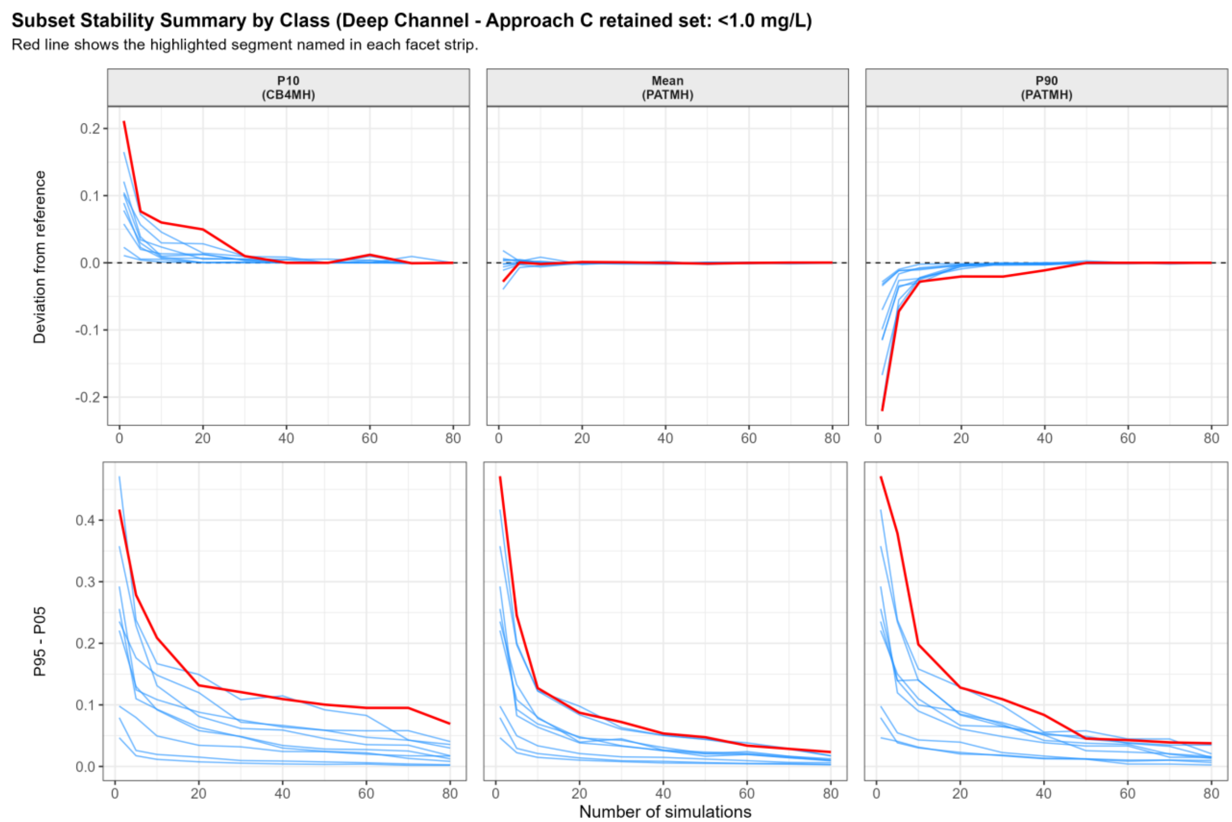

Figure #-2. Deep Channel retained-set subset-stability summary at the 1.0 mg/L threshold. The retained set contains all 10 Deep Channel segments. The upper row shows signed deviation from each segment's 100-simulation reference; the lower row shows the repeated-draw P95–P05 range.

Table #-3. Deep Channel stability summary (percentage points). Median and maximum stability measures across the 10 retained segments.

Stability measure	Num. Sim.	P10		Mean		P90	
		Median	Max. (segment)	Median	Max. (segment)	Median	Max. (segment)
Absolute deviation from reference	10	0.0117	0.0600 (CB4MH)	0.0013	0.0084 (CB4MH)	0.0222	0.0300 (CHSMH)
	20	0.0094	0.0496 (CB4MH)	0.0008	0.0022 (CB3MH)	0.0037	0.0204 (PATMH)
	40	0.0003	0.0084 (PATMH)	0.0005	0.0023 (CB4MH)	0.0013	0.0112 (PATMH)
	50	0.0000	0.0054 (CB3MH)	0.0007	0.0024 (CB4MH)	0.0000	0.0029 (CHSMH)
	60	0.0006	0.0119 (CB4MH)	0.0005	0.0015 (POTMH_MD)	0.0000	0.0015 (CB4MH)
	80	0.0000	0.0000 (all)	0.0002	0.0006 (POTMH_MD)	0.0000	0.0000 (all)
Repeated-draw P95-P05 range	10	0.101	0.209 (CB4MH)	0.074	0.127 (PATMH)	0.105	0.198 (PATMH)
	20	0.072	0.149 (PATMH)	0.043	0.098 (CB4MH)	0.075	0.129 (CB4MH)
	40	0.047	0.115 (PATMH)	0.026	0.054 (PATMH)	0.047	0.084 (PATMH)
	50	0.037	0.100 (CB4MH)	0.021	0.048 (PATMH)	0.035	0.058 (CB4MH)
	60	0.031	0.095 (CB4MH)	0.020	0.038 (POTMH_MD)	0.035	0.045 (CB4MH)
	80	0.017	0.069 (CB4MH)	0.010	0.023 (PATMH)	0.015	0.038 (PATMH)

Note: The median describes the typical segment; the maximum identifies the upper end of the segment results. The segment producing the maximum is shown in parentheses. Absolute deviations are rounded to four decimal places and repeated-draw ranges to three decimal places.

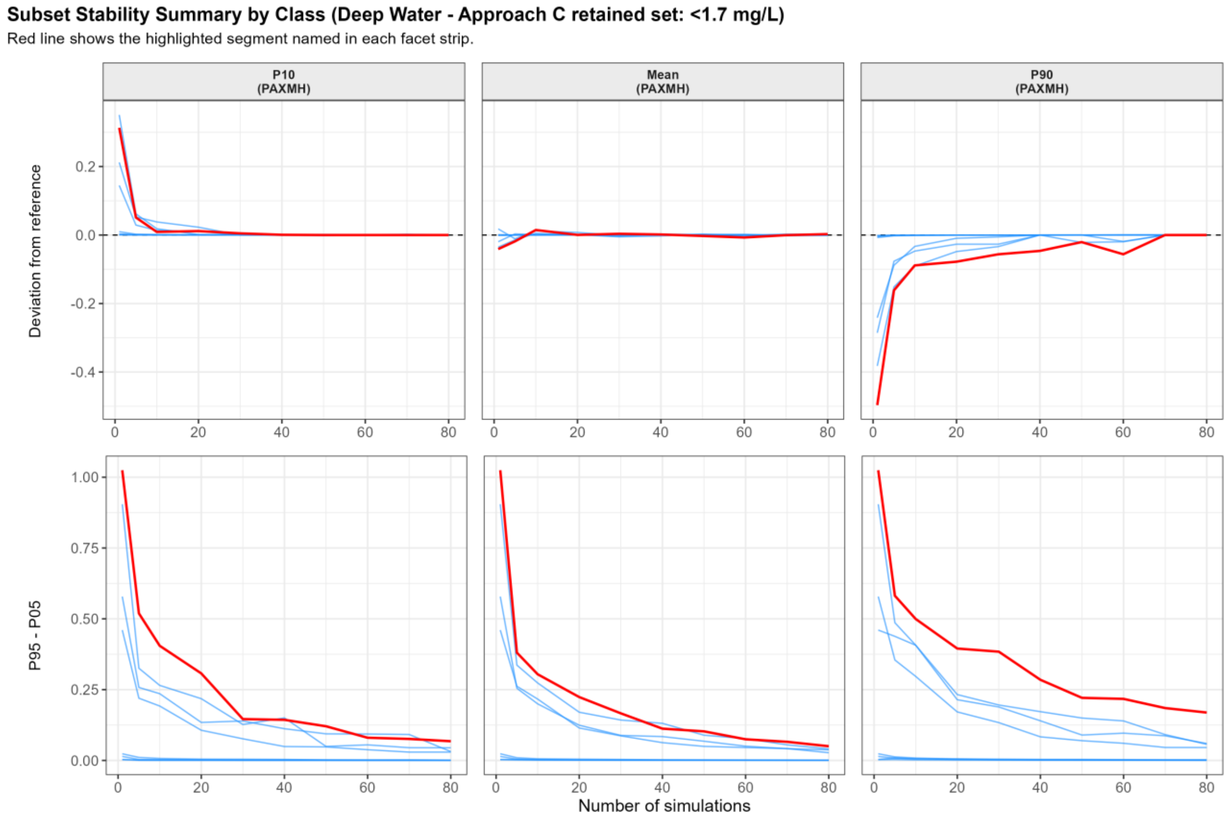

Figure #-3. Deep Water retained-set subset-stability summary at the 1.7 mg/L threshold. The retained set contains all 8 Deep Water segments. PAXMH is highlighted in red in each facet.

Table #-4. Deep Water stability summary (percentage points). Median and maximum stability measures across the 8 retained segments.

Stability measure	Num. Sim.	P10		Mean		P90	
		Median	Max. (segment)	Median	Max. (segment)	Median	Max. (segment)
Absolute deviation from reference	10	0.0052	0.0382 (MAGMH)	0.0009	0.0147 (PAXMH)	0.0172	0.0905 (SEVMH)
	20	0.0012	0.0229 (MAGMH)	0.0002	0.0076 (SEVMH)	0.0049	0.0779 (PAXMH)
	40	0.0001	0.0011 (SEVMH)	0.0009	0.0027 (SOUMH)	0.0000	0.0463 (PAXMH)
	50	0.0000	0.0017 (SEVMH)	0.0007	0.0032 (SEVMH)	0.0001	0.0216 (SEVMH)
	60	0.0000	0.0007 (CB7PH)	0.0004	0.0071 (PAXMH)	0.0004	0.0561 (PAXMH)
	80	0.0000	0.0000 (all)	0.0004	0.0027 (PAXMH)	0.0000	0.0000 (all)
Repeated-draw P95-P05 range	10	0.100	0.405 (PAXMH)	0.103	0.304 (PAXMH)	0.153	0.500 (PAXMH)
	20	0.056	0.307 (PAXMH)	0.060	0.224 (PAXMH)	0.088	0.395 (PAXMH)
	40	0.026	0.150 (SEVMH)	0.033	0.131 (SEVMH)	0.043	0.285 (PAXMH)
	50	0.025	0.121 (PAXMH)	0.026	0.103 (PAXMH)	0.037	0.221 (PAXMH)
	60	0.020	0.093 (MAGMH)	0.024	0.077 (SEVMH)	0.032	0.217 (PAXMH)
	80	0.015	0.068 (PAXMH)	0.014	0.050 (PAXMH)	0.024	0.169 (PAXMH)

Note: The median describes the typical segment; the maximum identifies the upper end of the segment results. The segment producing the maximum is shown in parentheses. Absolute deviations are rounded to four decimal places and repeated-draw ranges to three decimal places.

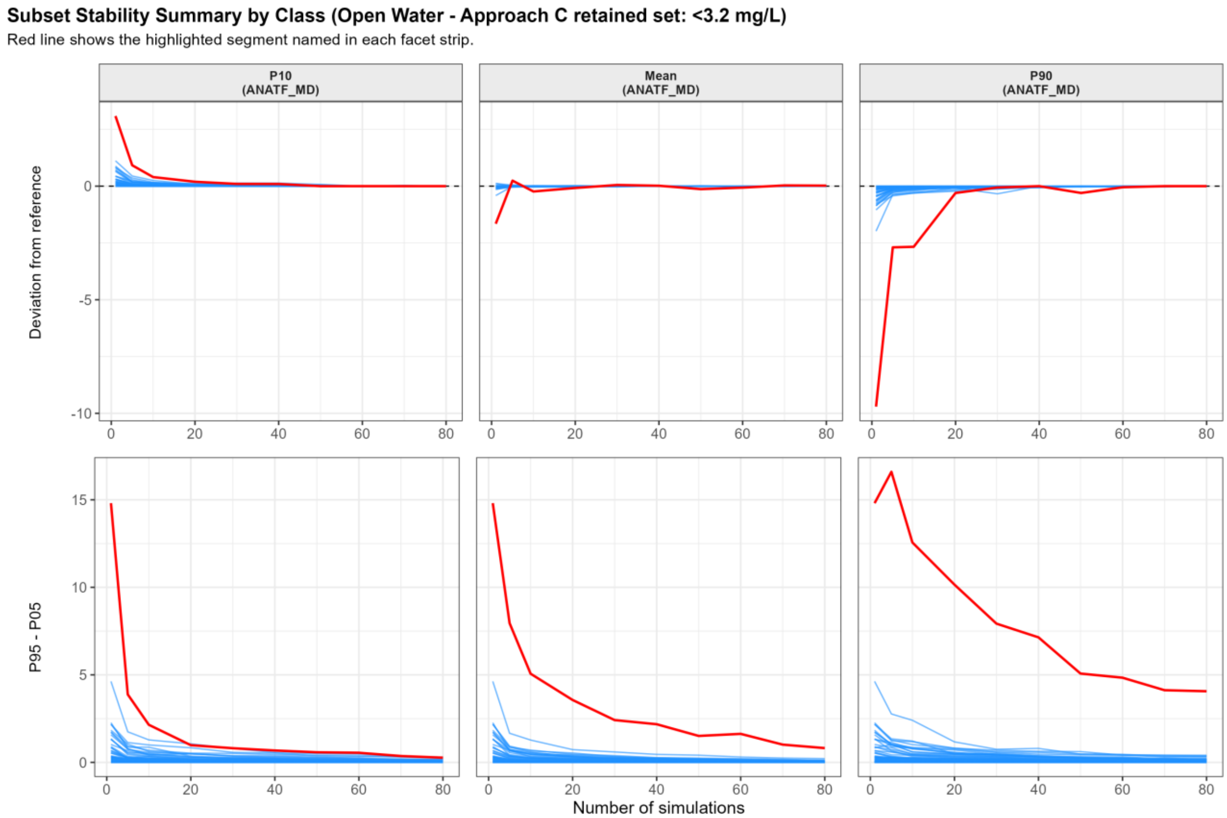

Figure #-4. Open Water retained-set subset-stability summary at the 3.2 mg/L threshold. The figure contains the 72 segments retained after excluding APPTF and NORTF. ANATF_MD is highlighted in red.

Table #-5. Open Water stability summary (percentage points). Median and maximum stability measures across the 72 retained segments.

Stability measure	Num. Sim.	P10		Mean		P90	
		Median	Max. (segment)	Median	Max. (segment)	Median	Max. (segment)
Absolute deviation from reference	10	0.0050	0.3993 (ANATF_MD)	0.0012	0.2352 (ANATF_MD)	0.0108	2.6692 (ANATF_MD)
	20	0.0023	0.1937 (ANATF_MD)	0.0008	0.0799 (ANATF_MD)	0.0047	0.3004 (ANATF_MD)
	40	0.0003	0.1344 (CHSTF)	0.0004	0.0181 (ANATF_MD)	0.0004	0.0709 (WBEMH)
	50	0.0000	0.0766 (CHSTF)	0.0004	0.1306 (ANATF_MD)	0.0001	0.3004 (ANATF_MD)
	60	0.0000	0.0160 (RHDMH)	0.0003	0.0649 (ANATF_MD)	0.0000	0.0470 (ANATF_MD)
	80	0.0000	0.0008 (CHOTF)	0.0002	0.0221 (ANATF_MD)	0.0000	0.0016 (LCHMH)
Repeated-draw P95-P05 range	10	0.051	2.148 (ANATF_MD)	0.064	5.065 (ANATF_MD)	0.106	12.561 (ANATF_MD)
	20	0.036	1.061 (CHSOH)	0.040	3.564 (ANATF_MD)	0.080	10.154 (ANATF_MD)
	40	0.016	0.679 (ANATF_MD)	0.025	2.179 (ANATF_MD)	0.054	7.141 (ANATF_MD)
	50	0.012	0.576 (ANATF_MD)	0.021	1.510 (ANATF_MD)	0.044	5.073 (ANATF_MD)
	60	0.011	0.548 (ANATF_MD)	0.017	1.628 (ANATF_MD)	0.045	4.837 (ANATF_MD)
	80	0.006	0.301 (CHSOH)	0.009	0.815 (ANATF_MD)	0.021	4.063 (ANATF_MD)

Note: The median describes the typical segment; the maximum identifies the upper end of the segment results. The segment producing the maximum is shown in parentheses. Absolute deviations are rounded to four decimal places and repeated-draw ranges to three decimal places.

#-3.4 Cross-class comparison

The retained analysis included all 10 Deep Channel analysis-class segments, all 8 Deep Water analysis-class segments, and 72 retained Open Water analysis-class segments. Open Water therefore accounted for 72 of the 90 segments included in the class-level results. It also showed the greatest variation among segments in the reference estimates calculated from all 100 simulations. **Table #-6** summarizes these reference-value ranges for each mutually exclusive analysis class.

Table #-6. Cross-class 100-simulation reference summary. Reference values are percentages of simulated grid-hours below the class threshold.

Designated-use class	Retained segments	Reference mean range (%)	Maximum reference P90 (%)
Deep Channel	10	0.025 to 4.707	4.892
Deep Water	8	0.001 to 2.755	3.155
Open Water	72	0.000002 to 7.017	13.073

Note: Open Water includes the 72 retained segments; APPTF and NORTF remain in the full inventory but are excluded from the stability summaries.

Across all three classes, the typical segment was generally well centered on the 100-simulation reference by moderate subset sizes. However, the repeated-draw P95–P05 ranges narrowed more gradually, particularly for P10 and P90. This distinction shows that an estimate can be centered near the reference while remaining sensitive to which simulations are included. **Table #-7** and **Table #-8** place the class medians in a common layout, allowing the typical response to be compared across the mutually exclusive analysis classes and simulation counts. Because the median does not describe the least-stable segments, these tables should be read together with the class-specific companion tables above. Those tables report the maximum value and identify the corresponding segment, providing a second view of how far the most sensitive segment differed from the typical class response.

Table #-7. Class-median absolute deviation from the 100-simulation reference by class, statistic, and selected simulation count.

Designated use class	Threshold (mg/L)	Statistic	Number of simulations					
			10	20	40	50	60	80
Deep Channel	1.0	P10	0.0117	0.0094	0.0003	<0.0001	0.0006	0.0000
Deep Channel	1.0	Mean	0.0013	0.0008	0.0005	0.0007	0.0005	0.0002
Deep Channel	1.0	P90	0.0222	0.0037	0.0013	0.0000	0.0000	0.0000
Deep Water	1.7	P10	0.0052	0.0012	<0.0001	<0.0001	<0.0001	0.0000
Deep Water	1.7	Mean	0.0009	0.0002	0.0009	0.0007	0.0004	0.0004
Deep Water	1.7	P90	0.0172	0.0049	<0.0001	0.0001	0.0004	0.0000
Open Water	3.2	P10	0.0050	0.0023	0.0003	<0.0001	0.0000	0.0000
Open Water	3.2	Mean	0.0012	0.0008	0.0004	0.0004	0.0003	0.0002
Open Water	3.2	P90	0.0108	0.0047	0.0004	0.0001	0.0000	0.0000

Table #-8. Class-median repeated-draw P95–P05 range by class, statistic, and selected simulation count.

Designated use class	Threshold (mg/L)	Statistic	Number of simulations					
			10	20	40	50	60	80
Deep Channel	1.0	P10	0.101	0.072	0.047	0.037	0.031	0.017
Deep Channel	1.0	Mean	0.074	0.043	0.026	0.021	0.020	0.010
Deep Channel	1.0	P90	0.105	0.075	0.047	0.035	0.035	0.015
Deep Water	1.7	P10	0.100	0.056	0.026	0.025	0.020	0.015
Deep Water	1.7	Mean	0.103	0.060	0.033	0.026	0.024	0.014
Deep Water	1.7	P90	0.153	0.088	0.043	0.037	0.032	0.024
Open Water	3.2	P10	0.051	0.036	0.016	0.012	0.011	0.006
Open Water	3.2	Mean	0.064	0.040	0.025	0.021	0.017	0.009
Open Water	3.2	P90	0.106	0.080	0.054	0.044	0.045	0.021

Note: Open Water entries summarize the 72 retained segments. APPTF and NORTF were excluded because their responses at the 3.2 mg/L threshold were trivial.

#-4. Synthesis

Three conclusions emerge from this analysis. First, increasing the simulation count generally improved reproducibility relative to the available 100-simulation reference. Second, the mean generally stabilized faster than P10 or P90. Third, the class median is useful for describing a typical segment, but the maximum is needed to identify segments for which a class-wide simulation count may provide less stability.

At 50 simulations, the class-median absolute deviations reported in **Table #-7** were less than 0.0001 percentage point for P10 in all three classes, 0.0004–0.0007 percentage points for the mean, and 0.0000–0.0001 percentage point for P90. The class-median repeated-draw ranges in **Table #-8** were 0.012–0.037, 0.021–0.026, and 0.035–0.044 percentage points, respectively. The class-specific maxima reported in **Table #-3** through **Table #-5** were much larger in a few cases: the largest absolute deviations across classes were 0.0766 for P10 (CHSTF), 0.1306 for the mean (ANATF_MD), and 0.3004 for P90 (ANATF_MD), while the largest repeated-draw ranges were 0.576, 1.510, and 5.073, all for ANATF_MD. **Table #-9** consolidates these n = 50 results.

Table #-9. Stability at 50 simulations by analysis class and statistic (percentage points). Cells report the class median followed by the class maximum and corresponding segment.

Analysis class	Statistic	Absolute deviation: median; maximum (segment)	P95–P05 range: median; maximum (segment)
Deep Channel	P10	<0.0001; 0.0054 (CB3MH)	0.037; 0.100 (CB4MH)
Deep Channel	Mean	0.0007; 0.0024 (CB4MH)	0.021; 0.048 (PATMH)
Deep Channel	P90	0.0000; 0.0029 (CHSMH)	0.035; 0.058 (CB4MH)
Deep Water	P10	<0.0001; 0.0017 (SEVMH)	0.025; 0.121 (PAXMH)
Deep Water	Mean	0.0007; 0.0032 (SEVMH)	0.026; 0.103 (PAXMH)
Deep Water	P90	0.0001; 0.0216 (SEVMH)	0.037; 0.221 (PAXMH)
Open Water	P10	<0.0001; 0.0766 (CHSTF)	0.012; 0.576 (ANATF_MD)
Open Water	Mean	0.0004; 0.1306 (ANATF_MD)	0.021; 1.510 (ANATF_MD)
Open Water	P90	0.0001; 0.3004 (ANATF_MD)	0.044; 5.073 (ANATF_MD)

Note: Values reproduce the n = 50 entries in Tables #-3 through #-5; class medians also correspond to Tables #-7 and #-8. Absolute deviations are shown to four decimal places and repeated-draw ranges to three decimal places.

The analysis was intended to identify diminishing returns rather than impose a numerical convergence tolerance after examining the results. For most retained segments, the central subset estimate was already close to the 100-simulation reference by approximately 40 simulations, and the additional narrowing from 50 to 60 simulations was generally modest. Taken together, the results support approximately 50–60 simulations as a reasonable provisional production range for most segments. Selecting 51 simulations also provides directly observed 6th-, 26th-, and 46th-order results for P10, the median, and P90 without interpolation. The n = 50 values in **Table #-9** quantify the reproducibility delivered near the lower end of that range and allow readers to evaluate it without treating the reported values as a formal pass-or-fail threshold.

ANATF_MD remains a documented high-sensitivity stress case rather than the basis for the standard ensemble size. Its n = 50 maxima in **Table #-9**, especially the 1.510-percentage-point mean range and 5.073-percentage-point P90 range, remain well above the corresponding class medians. ANATF_MD should therefore receive a prespecified segment-specific convergence check and, if necessary, additional simulations before its results are used in a final criteria determination.

A smaller ensemble of 21 simulations may be appropriate for testing and exploratory work when the objective is to verify workflow behavior rather than support a final criteria determination. This count also provides direct 3rd-, 11th-, and 19th-order results for P10, the median, and P90.

This analysis provides a defensible empirical basis for selecting simulation counts for testing, planning, and workflow development. Future sensitivity analyses could incorporate additional spatial measures and other indicators of simulation stability or assessment performance to help determine the final number of simulations used for assessment. The production count should also be evaluated using the CFD procedure and percentile-based decision rule ultimately adopted because they represent the final assessment endpoint.

#-5. Supporting exhibits

Exhibit #-1. Illustrative Monte Carlo Stability Analysis for One Deep-Channel Segment

This exhibit illustrates the logic used to evaluate how many simulations are needed to obtain stable results for a segment. The content is an example; its value is illustrative and is not used to calculate the updated class summaries in the main report.

The starting point is to treat each simulation as one replicate that reflects the typical variability and temporal/spatial associations estimated from observed data. For a given segment and dissolved oxygen threshold, each replicate is reduced to a single segment-level assessment endpoint: the percentage of simulated grid-hour values that fall below the threshold. In the example shown here, the threshold is 1.0 mg/L. When using 100 simulations, this produces a distribution of percent-below values, one value from each simulation, for a total of 100 values.

The normal probability plot shows the distribution of the percent-below values (**Figure #-5**). The key point is that the simulation output is not a single deterministic answer; it is a distribution of plausible answers induced by the simulation framework. For this example, the percent-below values have a mean of 4.771%, with P10 = 4.662%, P50 = 4.764%, and P90 = 4.924%. The probability plot indicates a roughly smooth, unimodal distribution with some departure from normality in the tails.

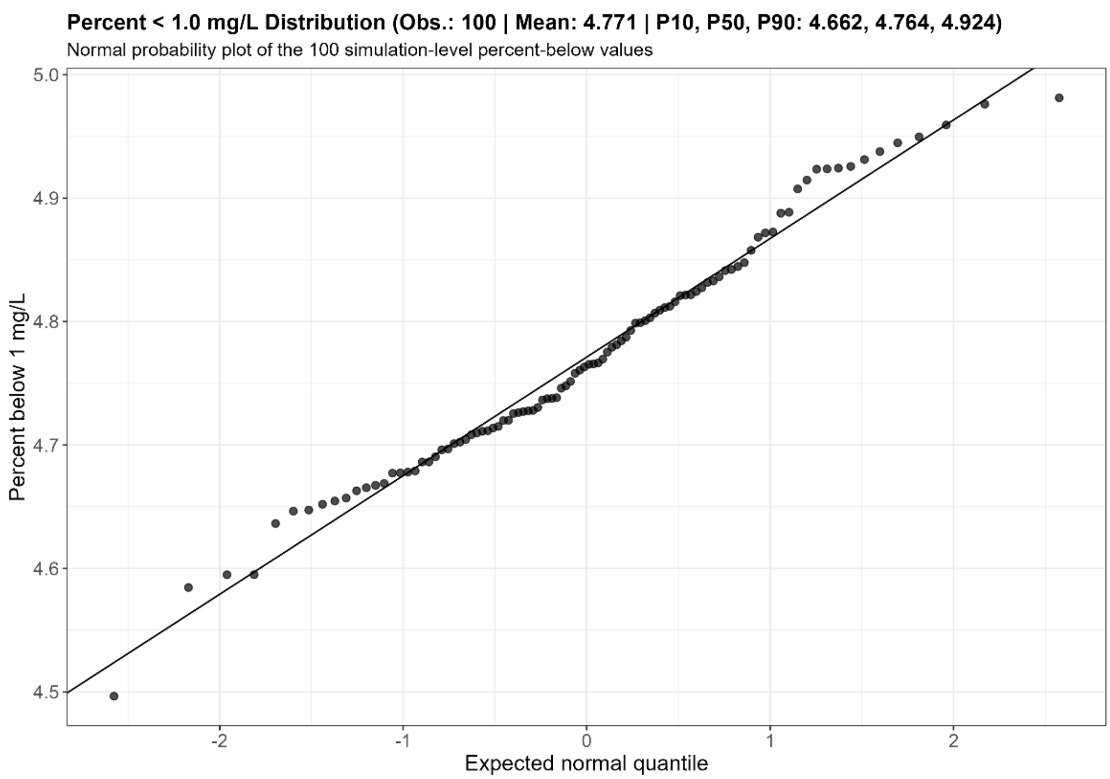

Figure #-5. Normal probability plot of the 100 simulation-level percent-below values for an illustrative deep-channel segment at the 1.0 mg/L threshold. Mean = 4.771%; P10 = 4.662%; P50 = 4.764%; P90 = 4.924%.

The density plot (**Figure #-6**) portrays the same percent-below values. Vertical lines identify P10, the mean, and P90, and the horizontal interval at the bottom shows the same summary range graphically. These statistics represent the lower tail, central tendency, and upper tail of the simulation-level distribution of the assessment endpoint values.

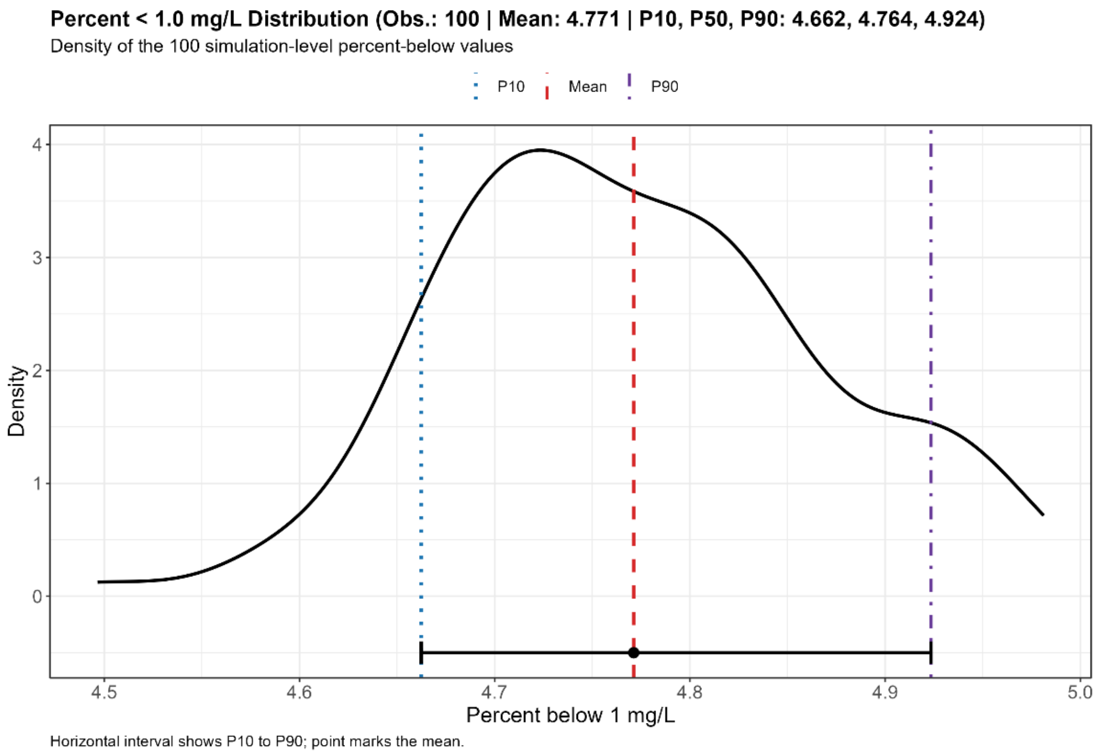

Figure #-6. Density plot of the 100 simulation-level percent-below values for an illustrative deep-channel segment at the 1.0 mg/L threshold. Vertical reference lines show P10 = 4.662%, Mean = 4.771%, and P90 = 4.924%; the horizontal interval spans P10 to P90.

The next step is to determine what would have happened if fewer than 100 simulations had been run. The 100 simulation-level values are treated as the reference set. A subset size n is selected, such as 5, 20, or 40, and n values are randomly sampled without replacement. P10, the mean, and P90 are calculated from each subset. Repeating this process produces a distribution of subset-based estimates for each statistic and subset size.

The repeated subset estimates are compared with the corresponding statistics from the full 100-simulation reference set. In this way, the analysis mimics whether a smaller number of distinct simulations would have recovered the result obtained from all 100 runs.

The repeated subset-draw distributions for n = 5, 20, and 40 in **Figure #-7** show two main patterns. First, the mean stabilizes faster than the tail summaries. Second, P10 and P90 are more sensitive to simulation count. At n = 5, the subset distributions are broad and may be displaced from the reference. By n = 20 and n = 40, the intervals become progressively narrower and better centered.

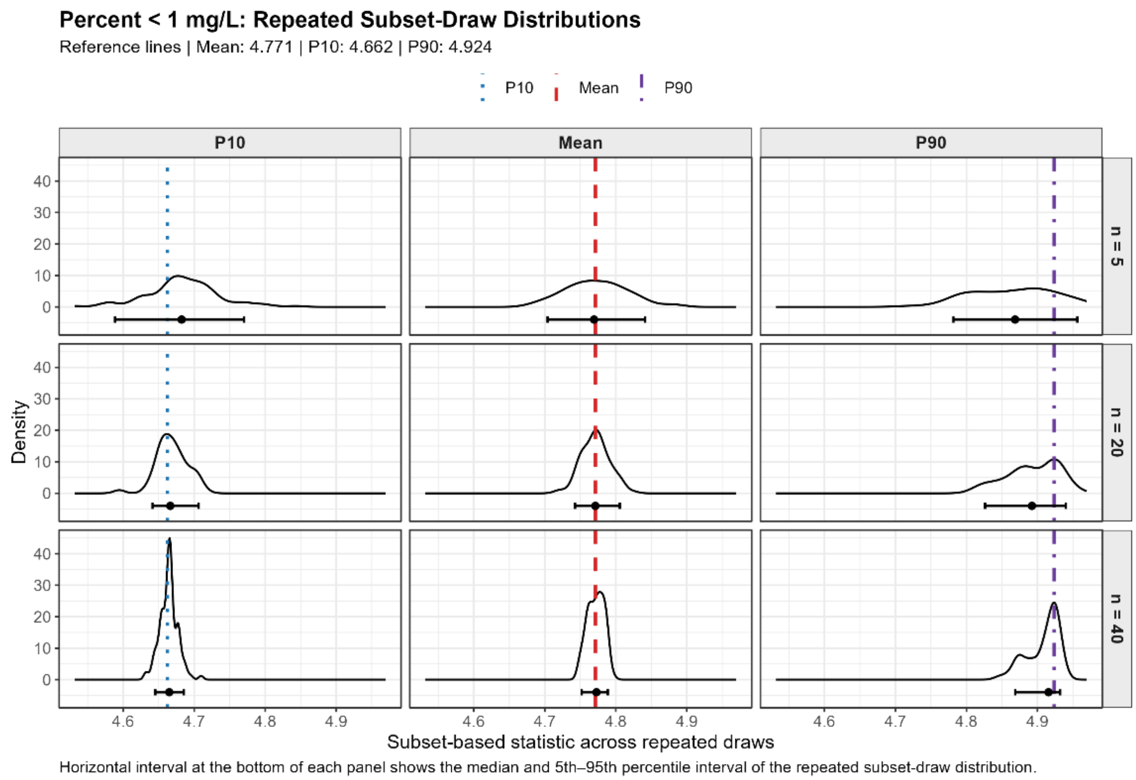

Figure #-7. Repeated subset-draw distributions for an illustrative deep-channel segment at the 1.0 mg/L threshold. Panels show P10, Mean, and P90 for subset sizes n = 5, 20, and 40 relative to the 100-simulation reference values.

The final figure (**Figure #-8**) extends the same calculation across n = 1, 5, 10, 20, 30, 40, 50, 60, 70, and 80. The points show the median repeated-draw estimate, vertical intervals show the 5th–95th percentile range, and dashed lines show the 100-simulation reference values.

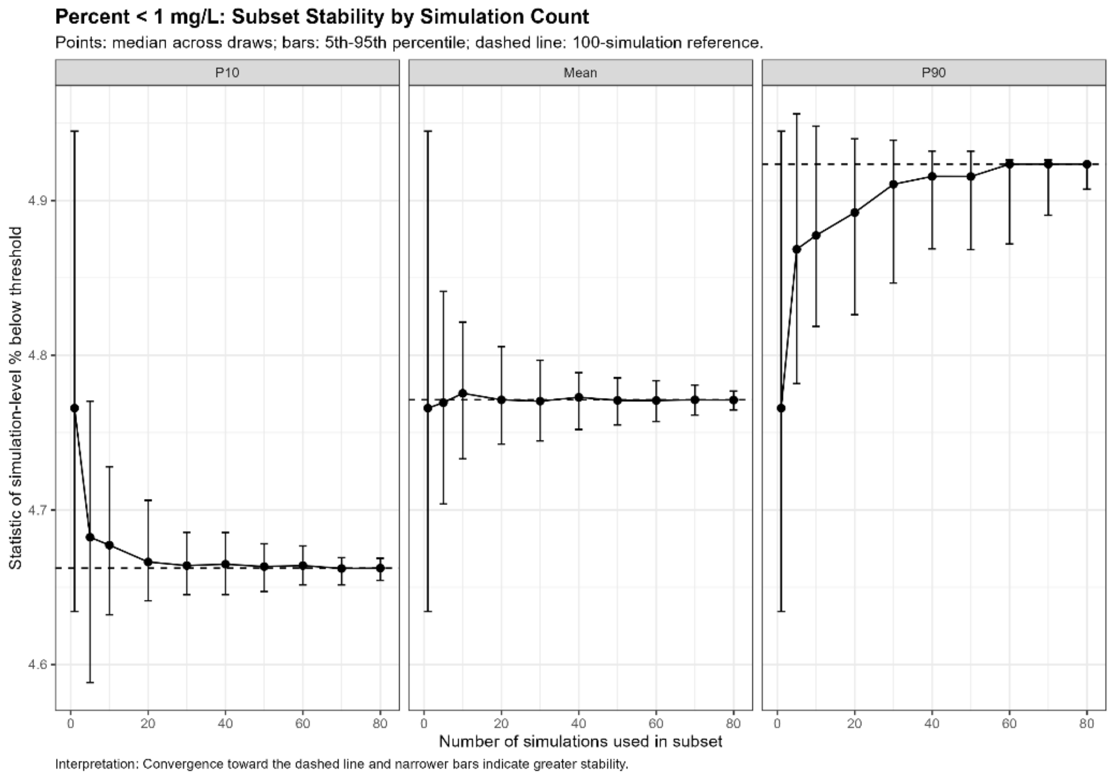

Figure #-8. Subset stability by simulation count for an illustrative deep-channel segment at the 1.0 mg/L threshold. P10, Mean, and P90 are shown for n = 1, 5, 10, 20, 30, 40, 50, 60, 70, and 80 relative to the 100-simulation reference.

Exhibit #-2. Investigation of ANATF_MD Simulation-Count Sensitivity

ANATF_MD was the dominant Open Water exception in the subset-stability results. Its subset medians approached the 100-simulation reference as the number of simulations increased, but the repeated-draw ranges remained much wider than those of most other retained Open Water segments, particularly for P90 (**Figure #-9**).

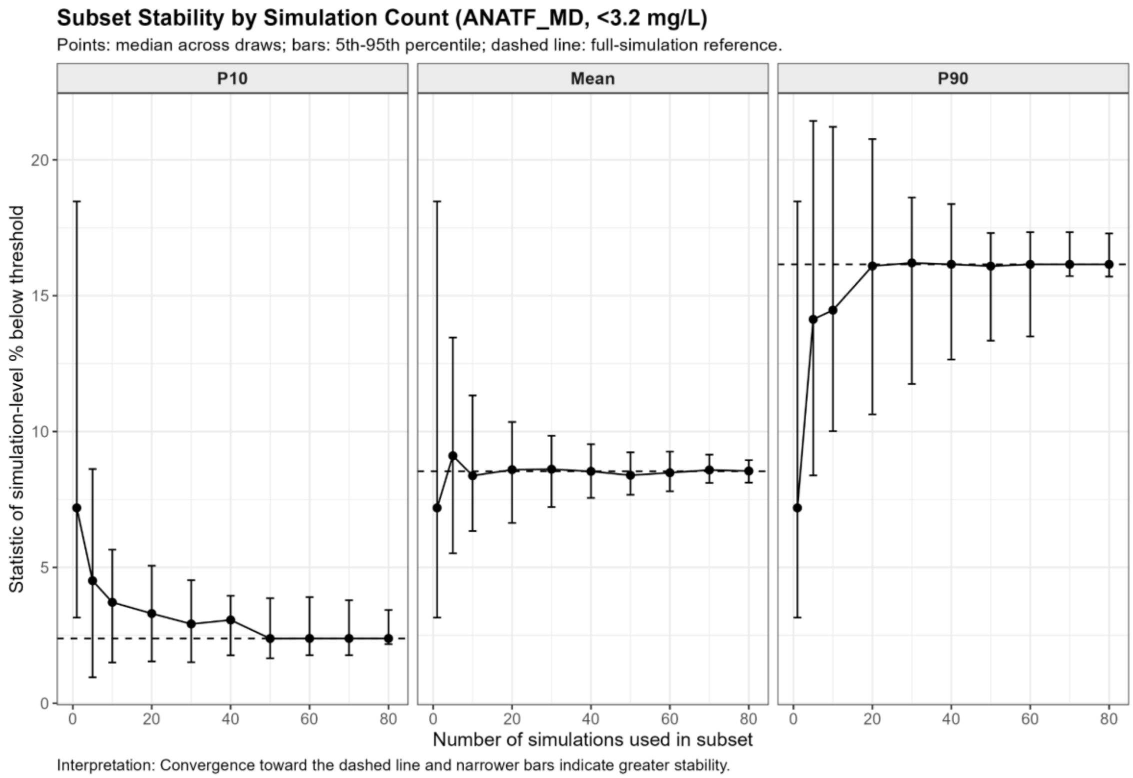

Figure #-9. Baseline ANATF_MD subset stability at the 3.2 mg/L threshold. Points show repeated-draw medians, bars show the 5th–95th percentile range, and dashed lines show the 100-simulation reference values.

The interpolated mean-model distribution (**Figure #-10**) provides a direct explanation for this sensitivity. ANATF_MD places an unusually large share of its grid-hour values close to the 3.2 mg/L threshold. Small, scientifically plausible residual changes can therefore move many spatially and temporally correlated grid-hours across the threshold together.

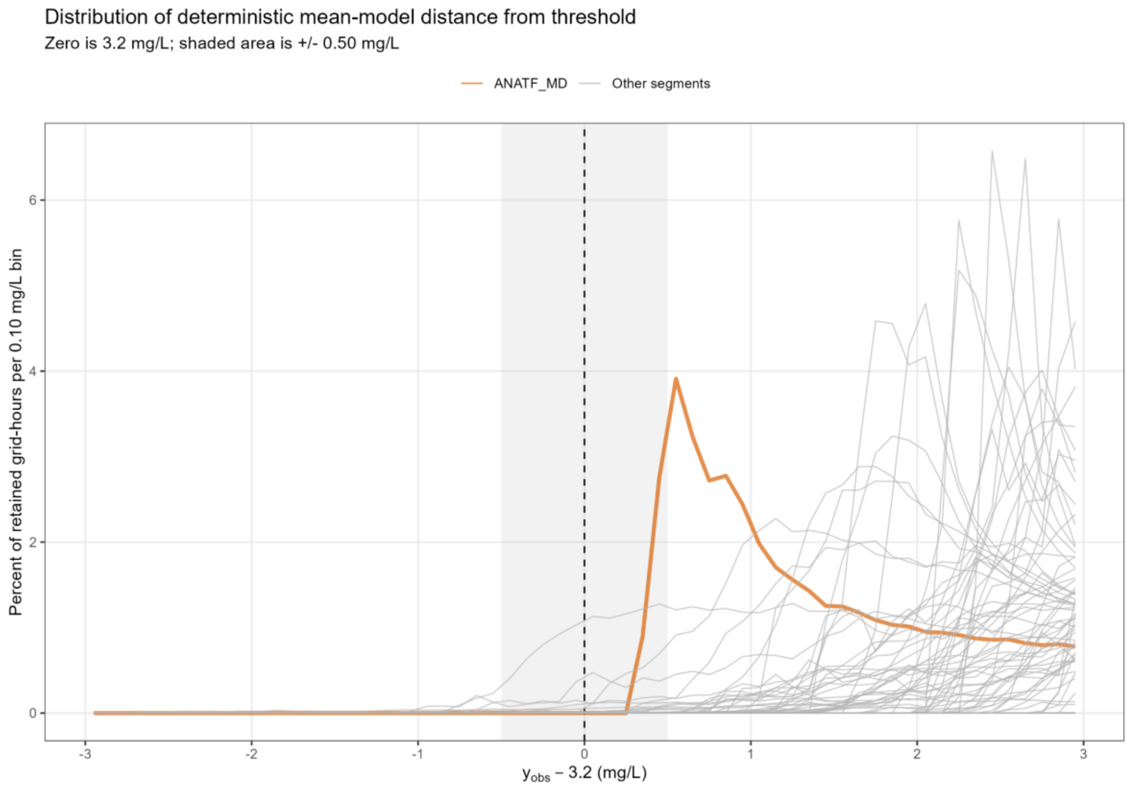

Figure #-10. Distribution of deterministic mean-model distance from the 3.2 mg/L threshold. The orange line identifies ANATF_MD and gray lines identify the other Open Water segments. Zero corresponds to 3.2 mg/L, and the shaded region spans ±0.50 mg/L.

Table #-10. ANATF_MD threshold-proximity diagnostics.

Diagnostic	ANATF_MD result	Comparative interpretation
Retained grid-hours	25,185	4th-smallest of 74
Minimum y_obs	3.543 mg/L	0.343 mg/L above threshold
5th percentile y_obs	3.738 mg/L	Lower tail remains close to threshold
10th percentile y_obs	3.872 mg/L	0.672 mg/L above threshold
Within ±0.50 mg/L	3.66%	Rank 2 of 74
Within ±1.00 mg/L	18.74%	Rank 1 of 74
Residual sigma	0.496 mg/L	Moderate rather than class-extreme

ANATF_MD ranked second among the 74 Open Water segments for the percentage of deterministic predictions within ±0.50 mg/L of 3.2 mg/L and first within ±1.00 mg/L. Its residual sigma of approximately 0.496 mg/L was elevated but not an extreme class outlier. Additional diagnostic analyses were conducted to investigate whether alternative correlation and residual-variation settings could account for the ANATF_MD behavior. The combination of threshold proximity, residual amplitude, comparatively limited grid-hour support, and strong correlation provides a more complete explanation than any one parameter considered alone. These exploratory analyses did not identify any production configuration changes and therefore are not presented in this report.

The practical implication is that ANATF_MD should be retained as a documented high-sensitivity case and subjected to a prespecified segment-specific convergence check. Its small segment grid, combined with a large proportion of space-time grid-hours near the threshold, creates genuine uncertainty in the stability of its assessment status. If the provisional production ensemble does not produce a stable result, additional simulations should be evaluated; if the result remains unstable, the assessment should be explicitly qualified. This segment-specific treatment prevents ANATF_MD from determining the standard ensemble size for the generally more stable segments.