

4D Interpolator

Number of Simulations

Bay Oxygen Research Group
August 17, 2026

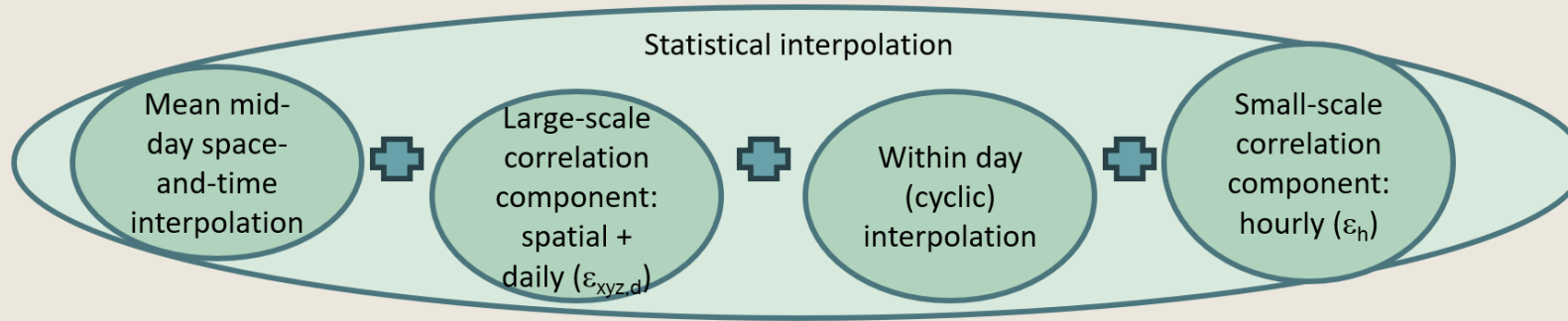
Jon Harcum¹ for the development team:
Elgin Perry², Rebecca Murphy³, Breck Sullivan⁴, Peter Tango⁴

¹ Tetra Tech, ² Statistics Consultant, ³ UMCES at CBP, ⁴ USGS at the CBP

Outline

- Brief review: building realistic simulated DO.
- From one simulation to an ensemble of simulated outcomes.
- Converting each simulation to a segment-level assessment endpoint.
- Testing endpoint stability as simulation count increases.
- Implications for testing and production simulation counts.

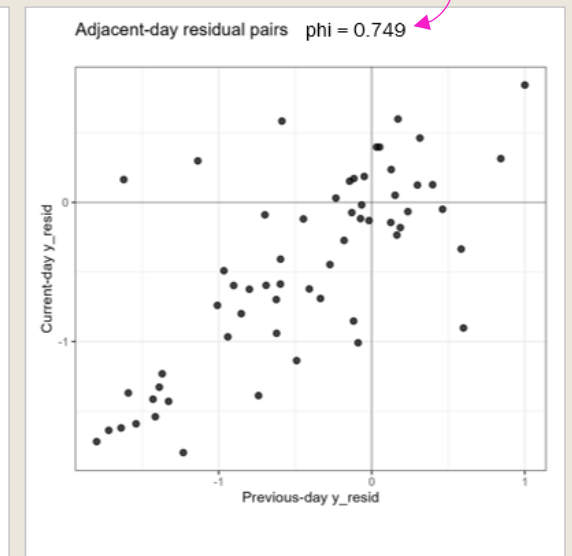
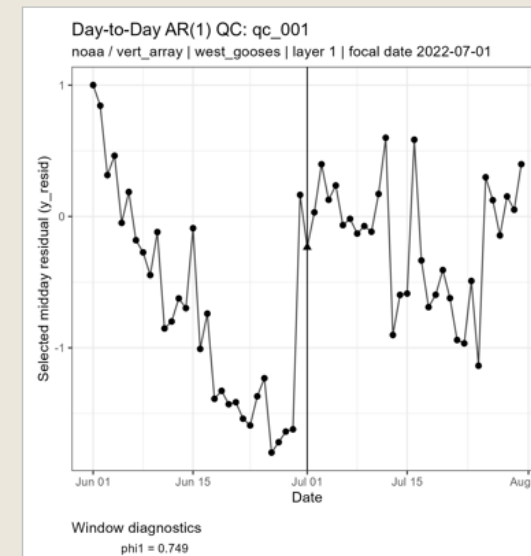
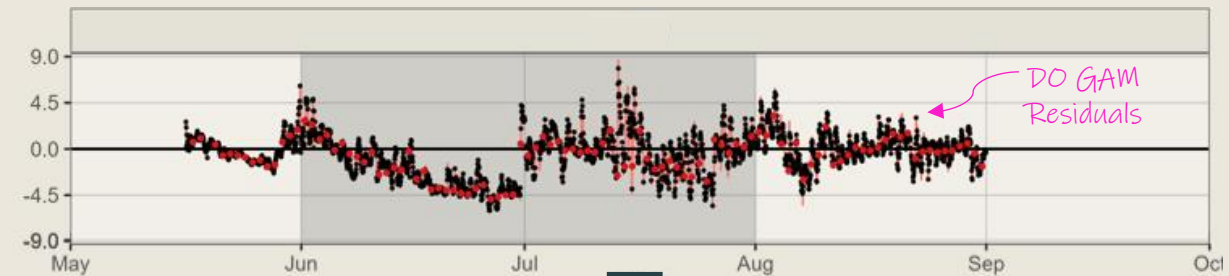
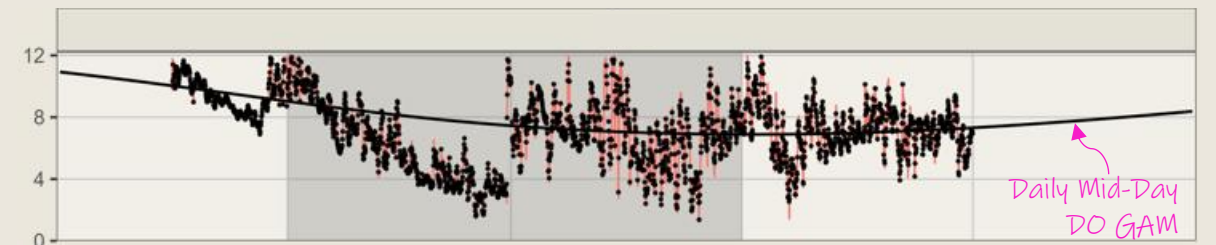
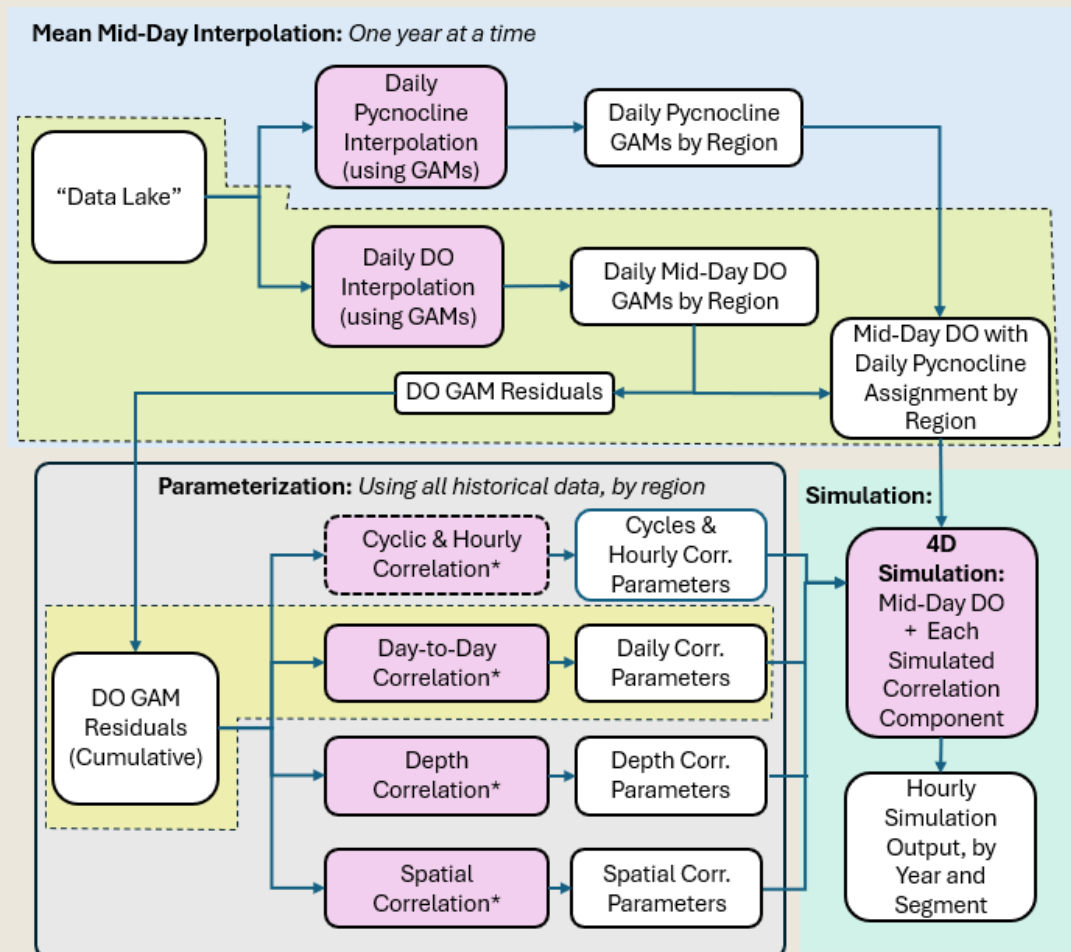
Building Realistic DO Simulations in Components



- Begin with the **expected mid-day DO pattern** across season, location, and depth.
- Add **large-scale correlated departures** from that pattern:
 - *day-to-day correlation*
 - *depth correlation*
 - *spatial correlation*
- Then add the **within-day cycle** and **hourly correlated variation**.

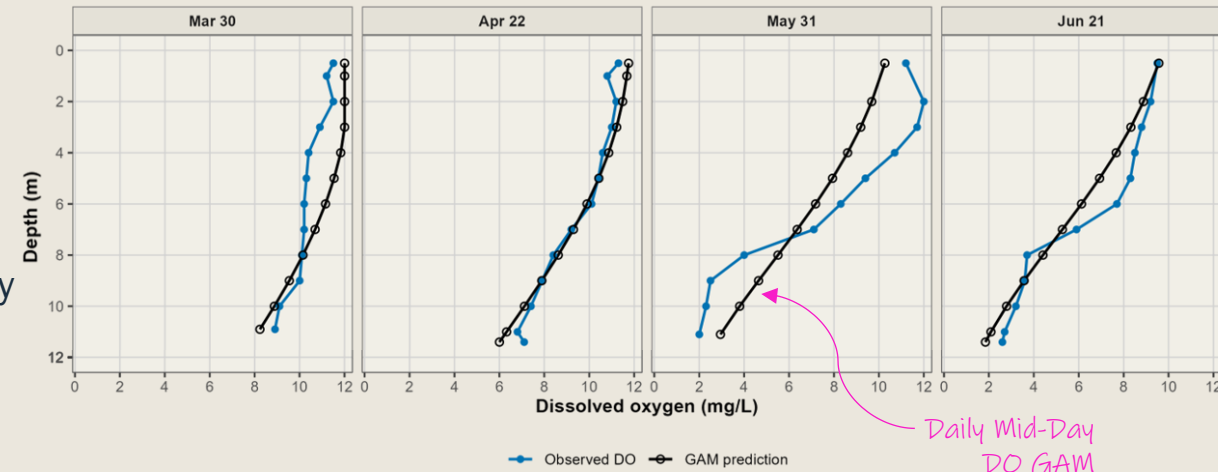
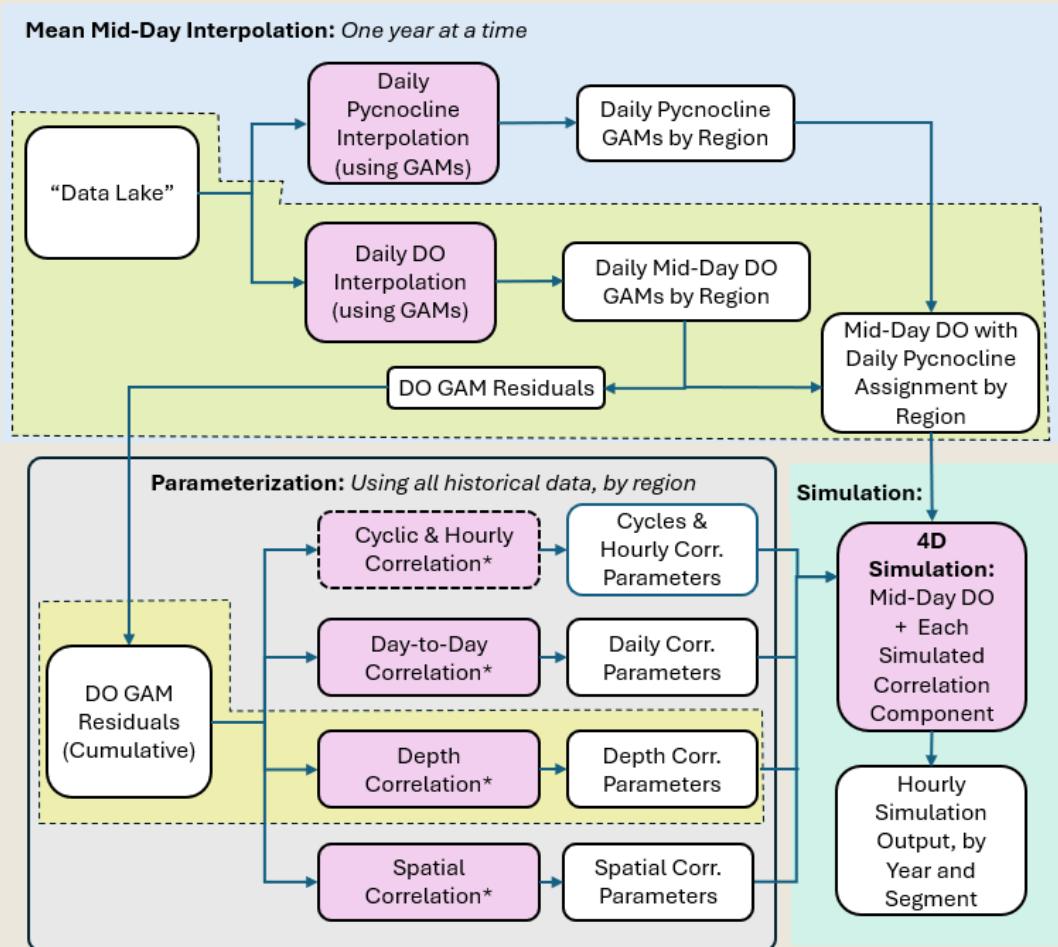
Day-to-Day Correlation Preserves Temporal Persistence

- The mid-day GAM provides the expected DO pattern.
- Residuals measure departures from that expectation.
- Data show positive day-to-day correlation which allows the departures to persist into the following day.



Depth Correlation Preserves Vertical Coherence

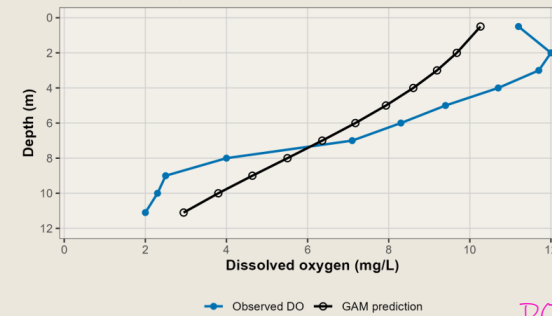
- The mid-day GAM provides the expected vertical DO profile.
- Residuals describe departures from that profile at each depth.
- Data show positive depth correlation making departures at nearby depths move together without making the depths identical.



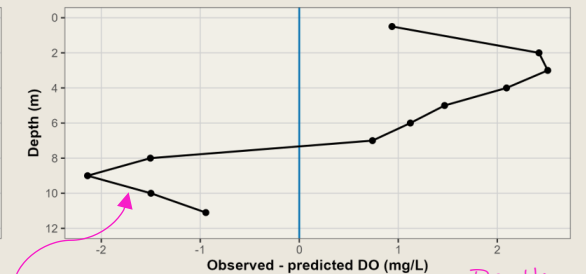
XE3551 worked profile: May 31, 2022

Direct spacing = 1 m; retained depths = 10; adjacent pairs = 9; standardized 1-m correlation = 0.881

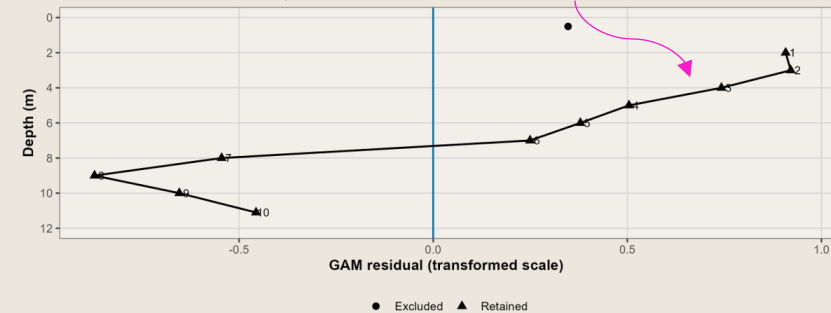
A. Observed and expected DO



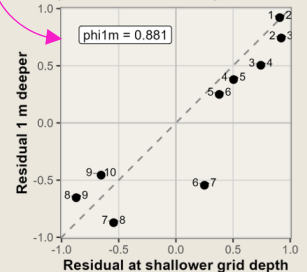
B. Residual in observed units



C. Model-scale residual and 1-m grid



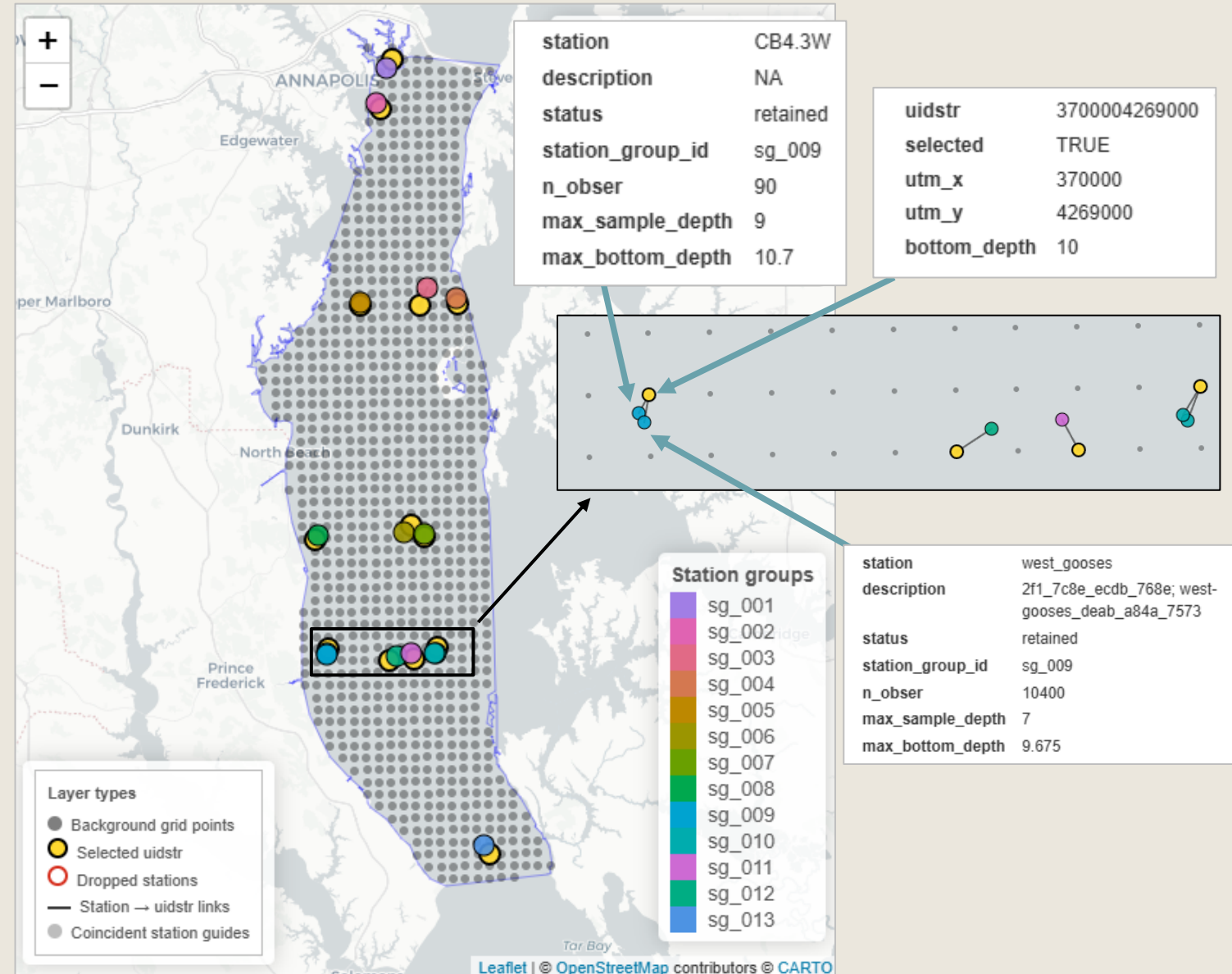
D. Adjacent 1-m residual pairs



Example Simulation Outputs

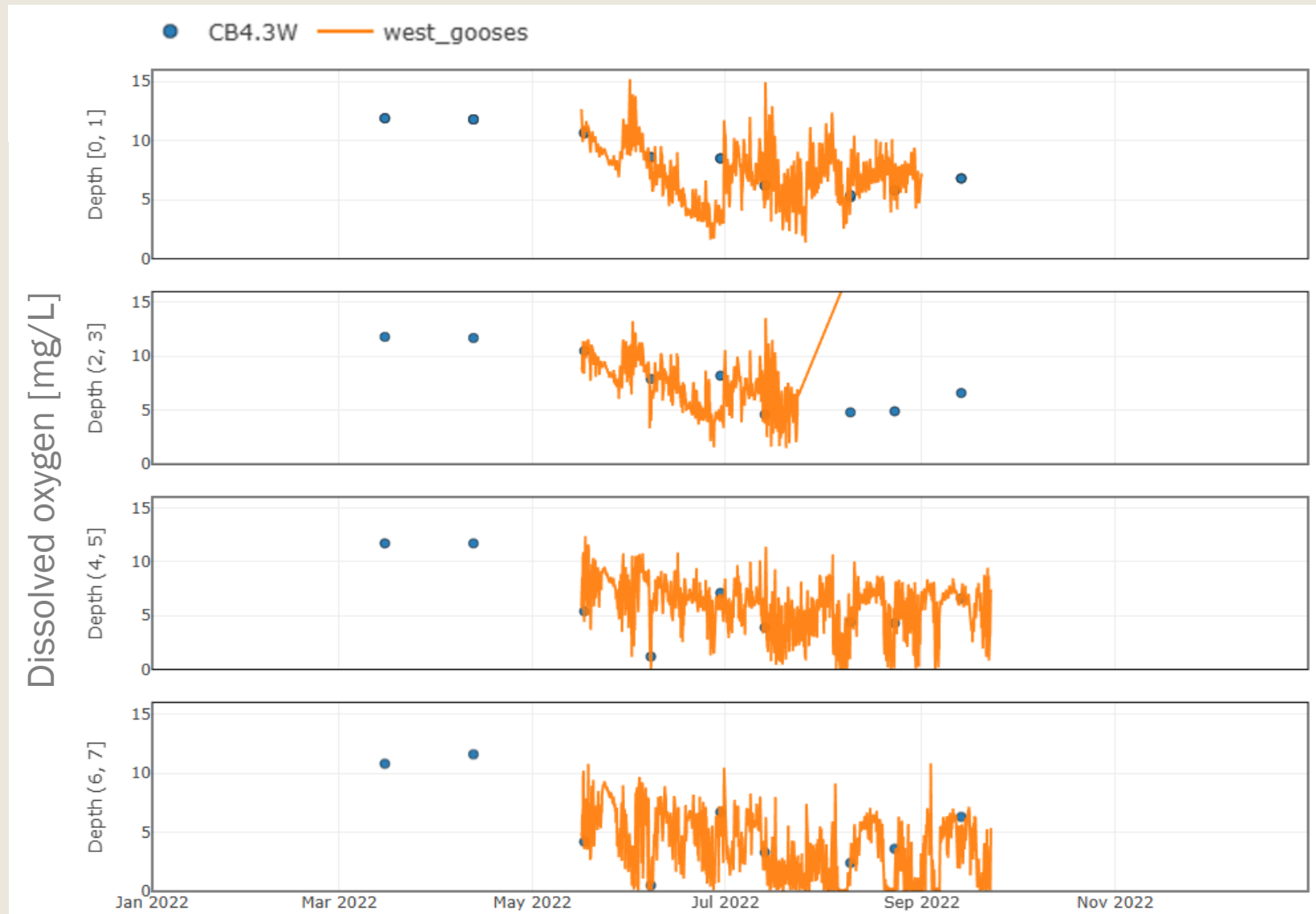
Simulate DO throughout the 4D grid

- CB4MH is represented by 1-km horizontal grid columns with multiple depth layers.
 - 2D: 886 grid points
 - 3D: 9,237 grid points
- Monitoring stations are grouped with nearby grid locations for visualization.
- The next slides follow one selected grid column and four depth bands.



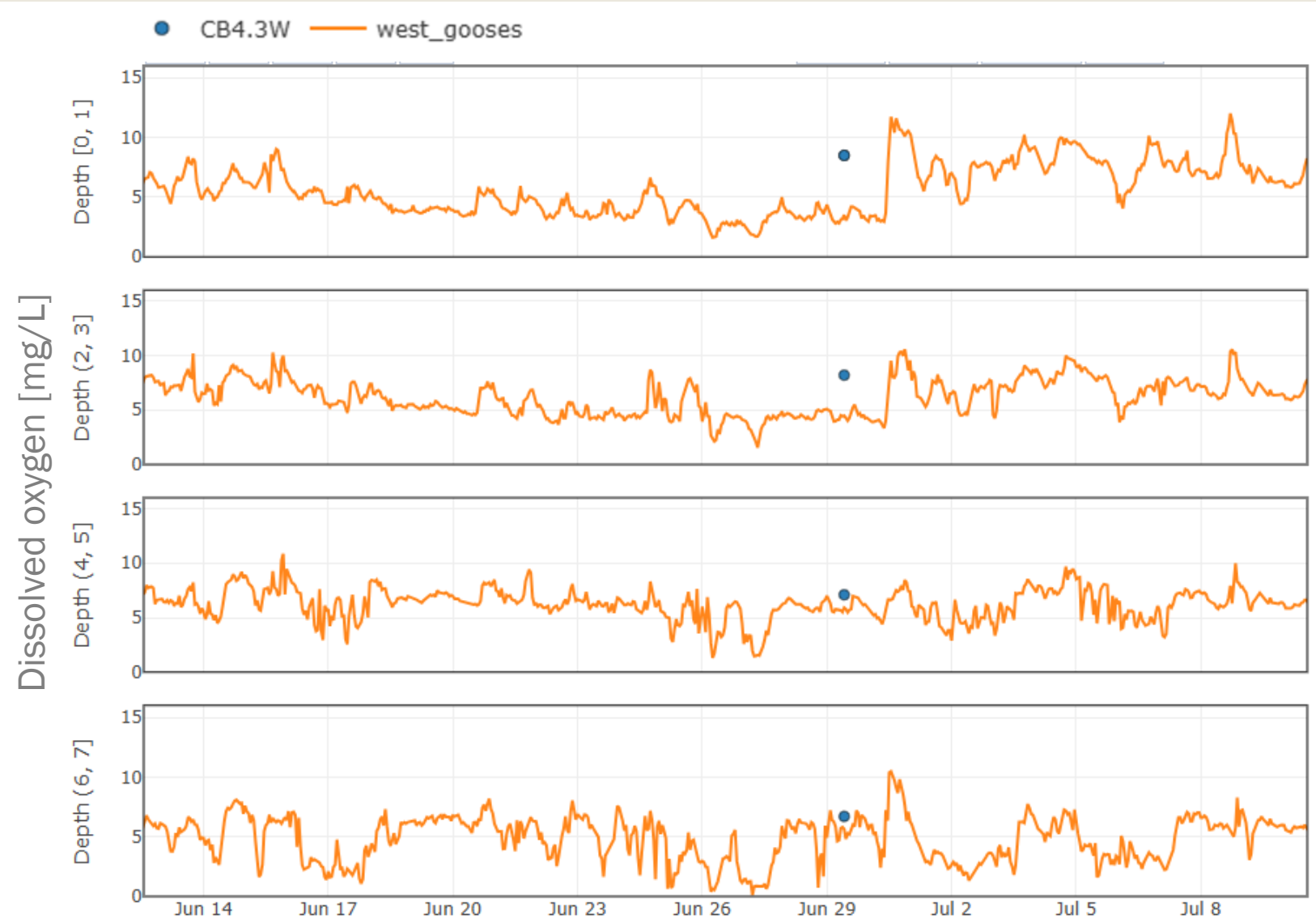
Monitoring data provide incomplete coverage

- The station group combines continuous monitoring records with discrete monitoring observations.
- Coverage differs across dates and depth bands.
- The model uses these observations to generate a complete simulated DO field.



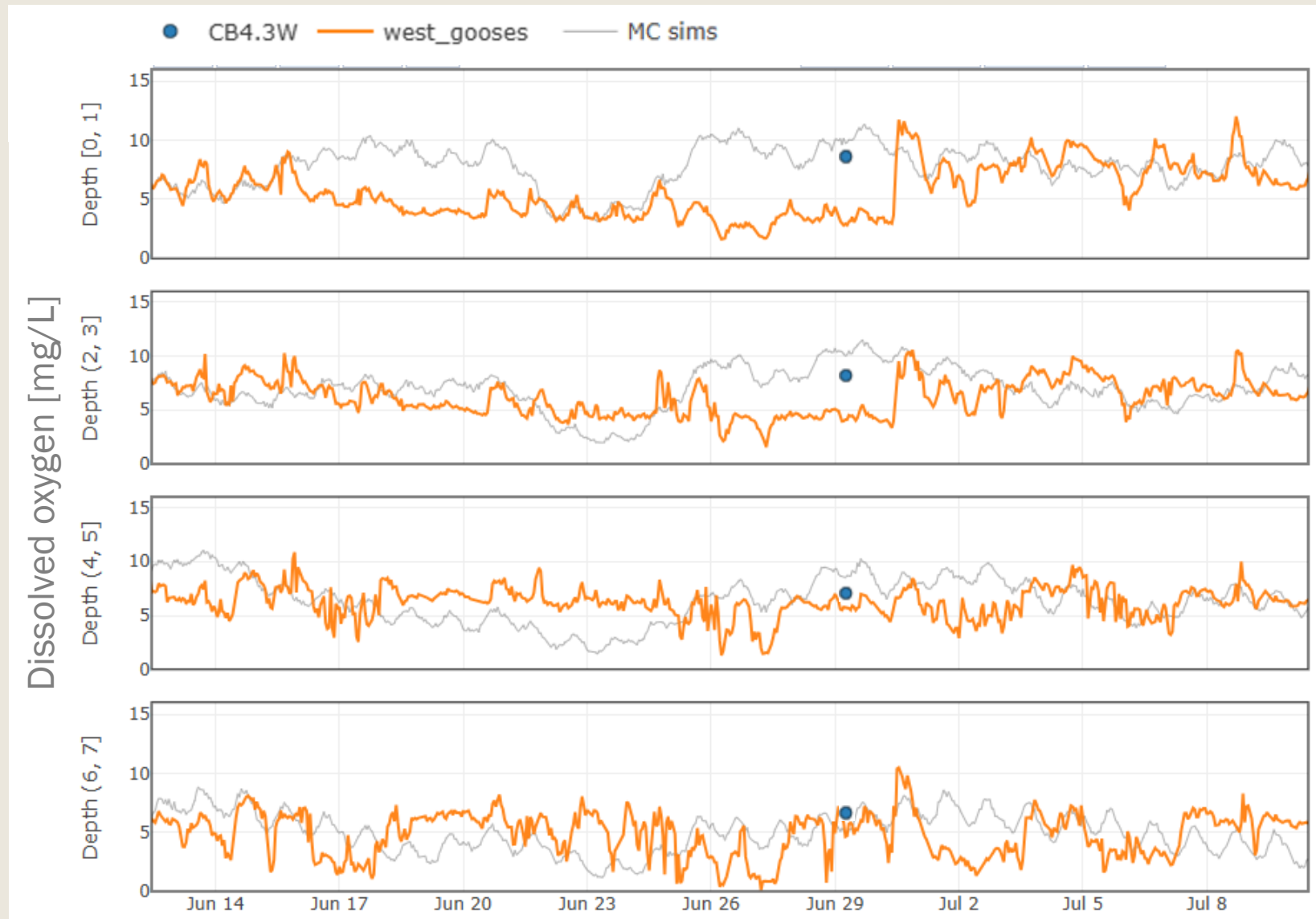
A closer view reveals temporal and vertical structure

- Nearby depth bands commonly change in the same direction, but their DO levels and excursions remain different.
- The record includes both gradual changes and short-lived departures.
- These are the temporal and vertical patterns that the simulation components are intended to reproduce.



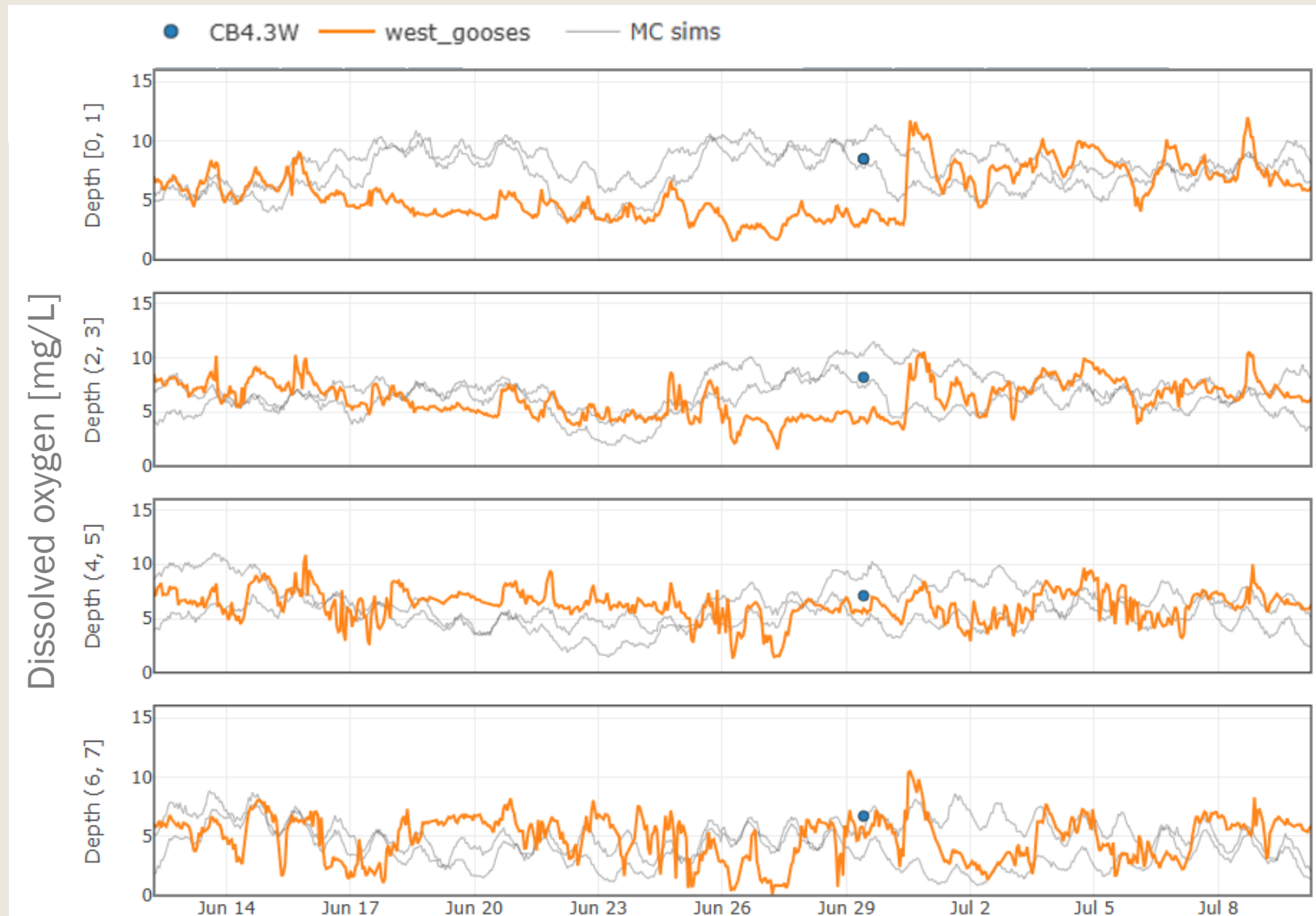
One simulation provides one complete simulated history

- The gray line is one simulation; the orange and blue marks show monitoring observations.
- The simulation combines the expected DO pattern with correlated departures.
- It represents the fitted structure (i.e., behavior) rather than attempting to reproduce every observation point-for-point.



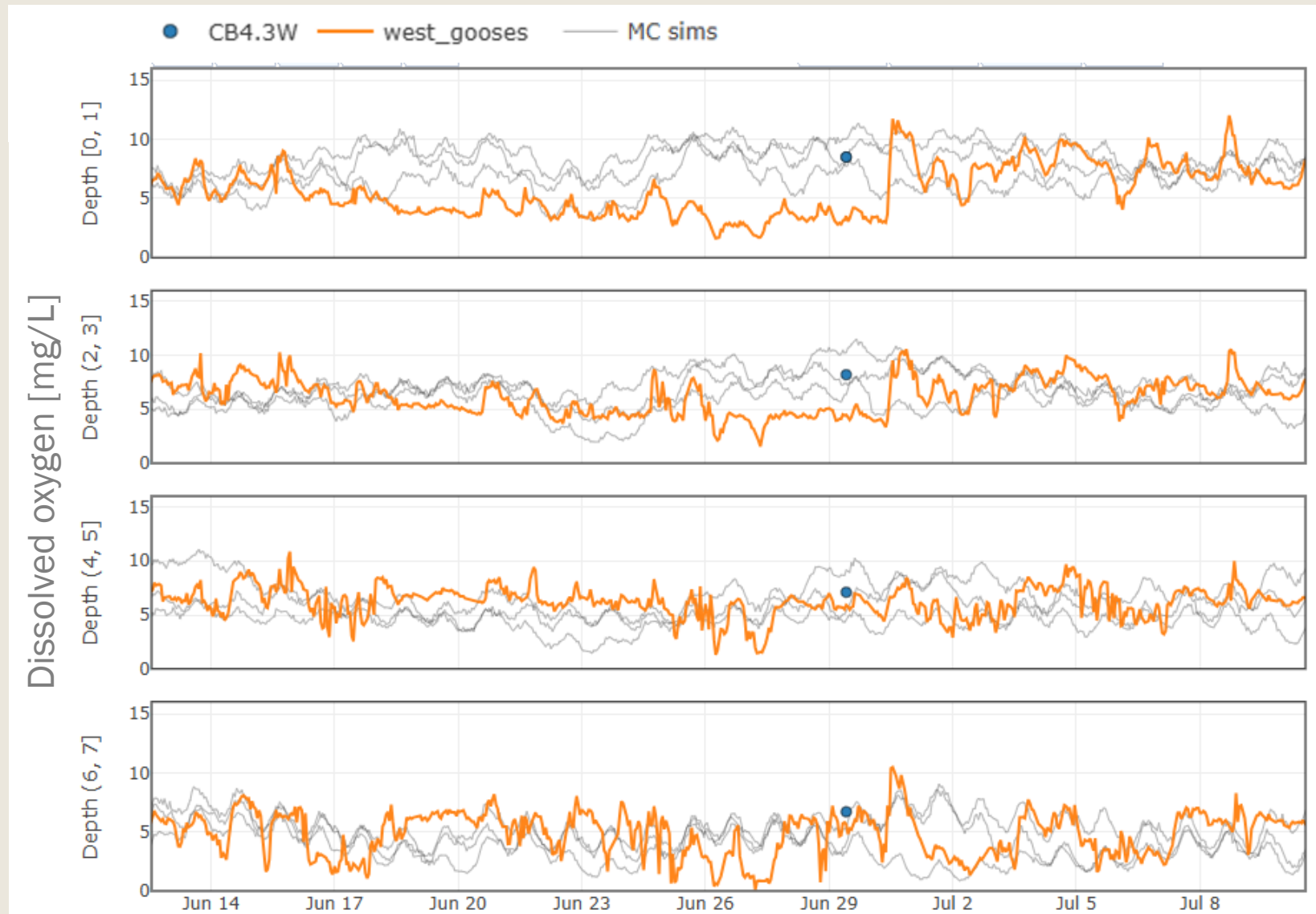
A second simulation produces a different history

- The expected pattern and fitted parameterization remain unchanged.
- Newly drawn correlated departures change the timing and magnitude of DO excursions.
- Neither simulation is the preferred trajectory; both represent expected variability.



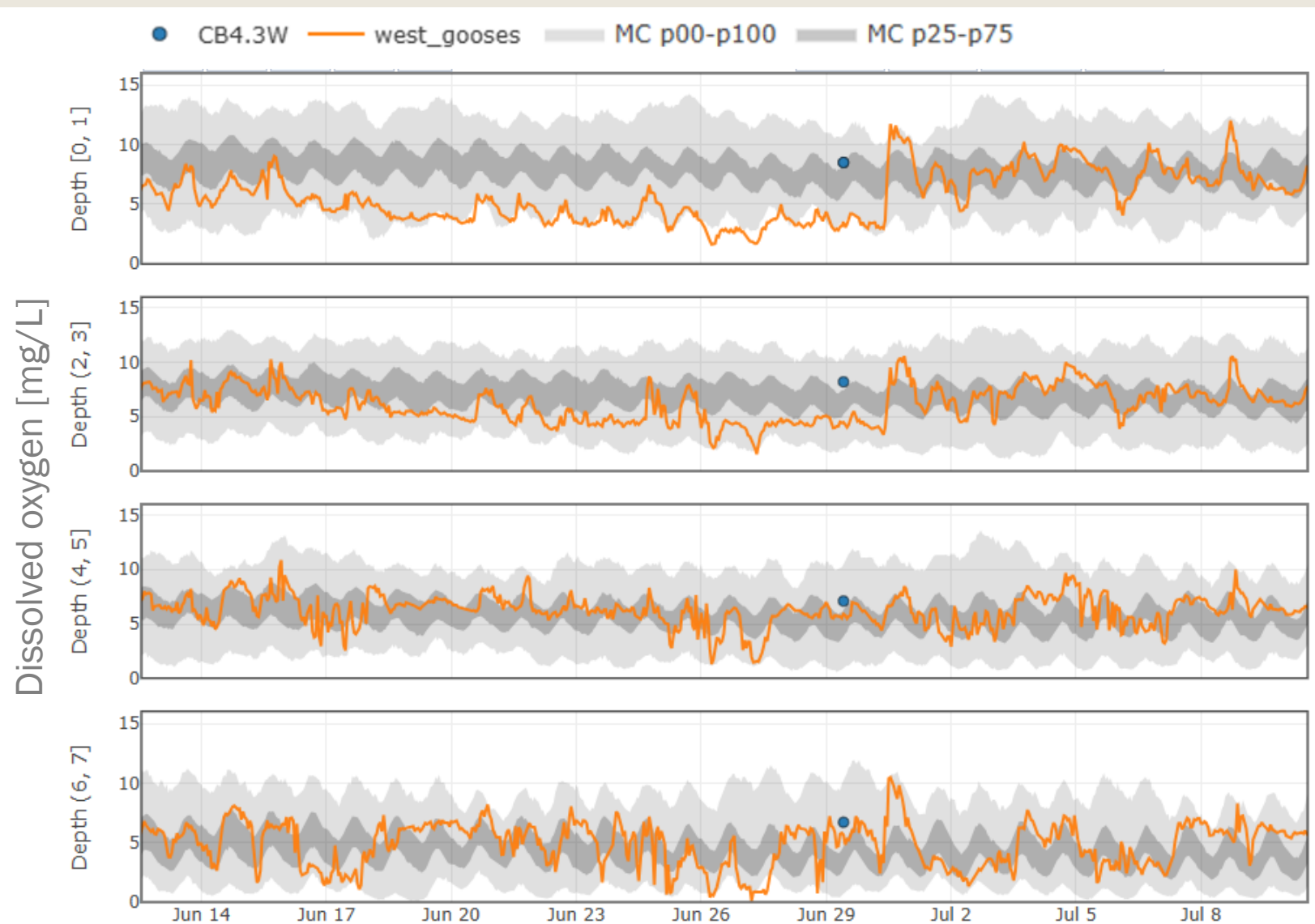
Additional simulations fill out the range of outcomes

- Adding simulations samples the same fitted model more completely.
- The center and range of simulated behavior begin to emerge.
- Tail behavior remains more sensitive to how many simulations have been generated.



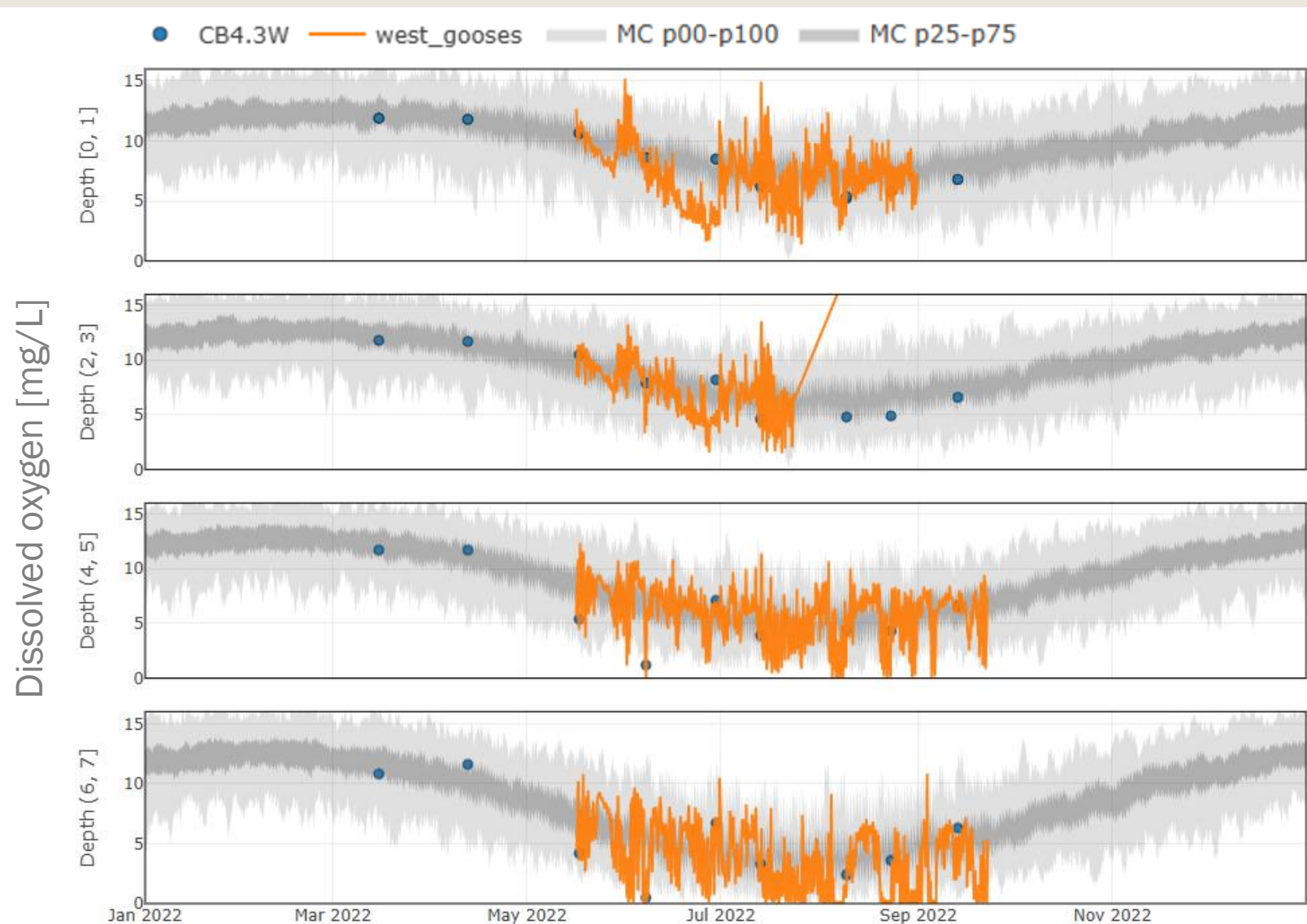
The ensemble summarizes variation across 100 simulations

- The dark band shows the central 50% of simulated values; the light band shows the full simulated range.
- The bands summarize results across simulations; neither band represents an individual simulation.



Each simulation represents one full year of hourly DO

- Assessment endpoints are calculated separately for each simulation, producing 100 endpoint values.

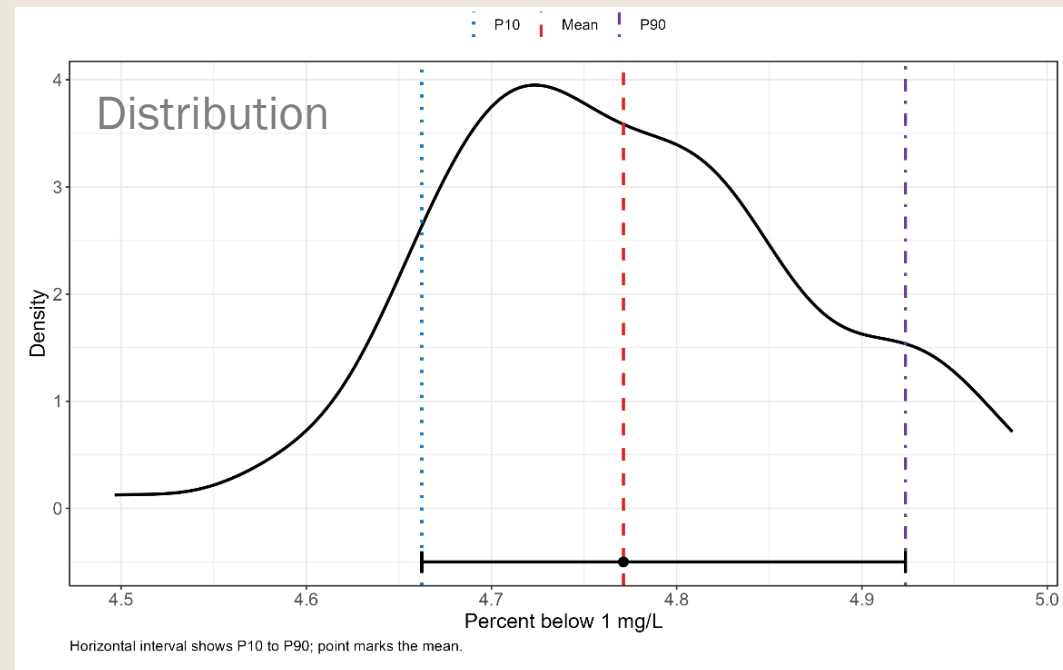
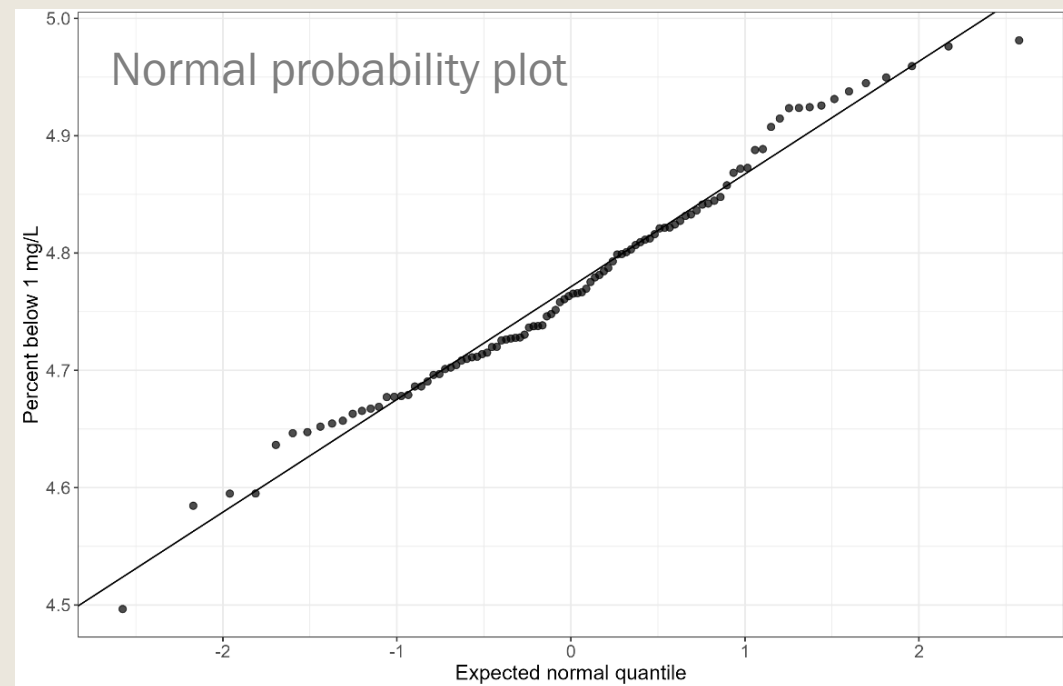


Number of simulations

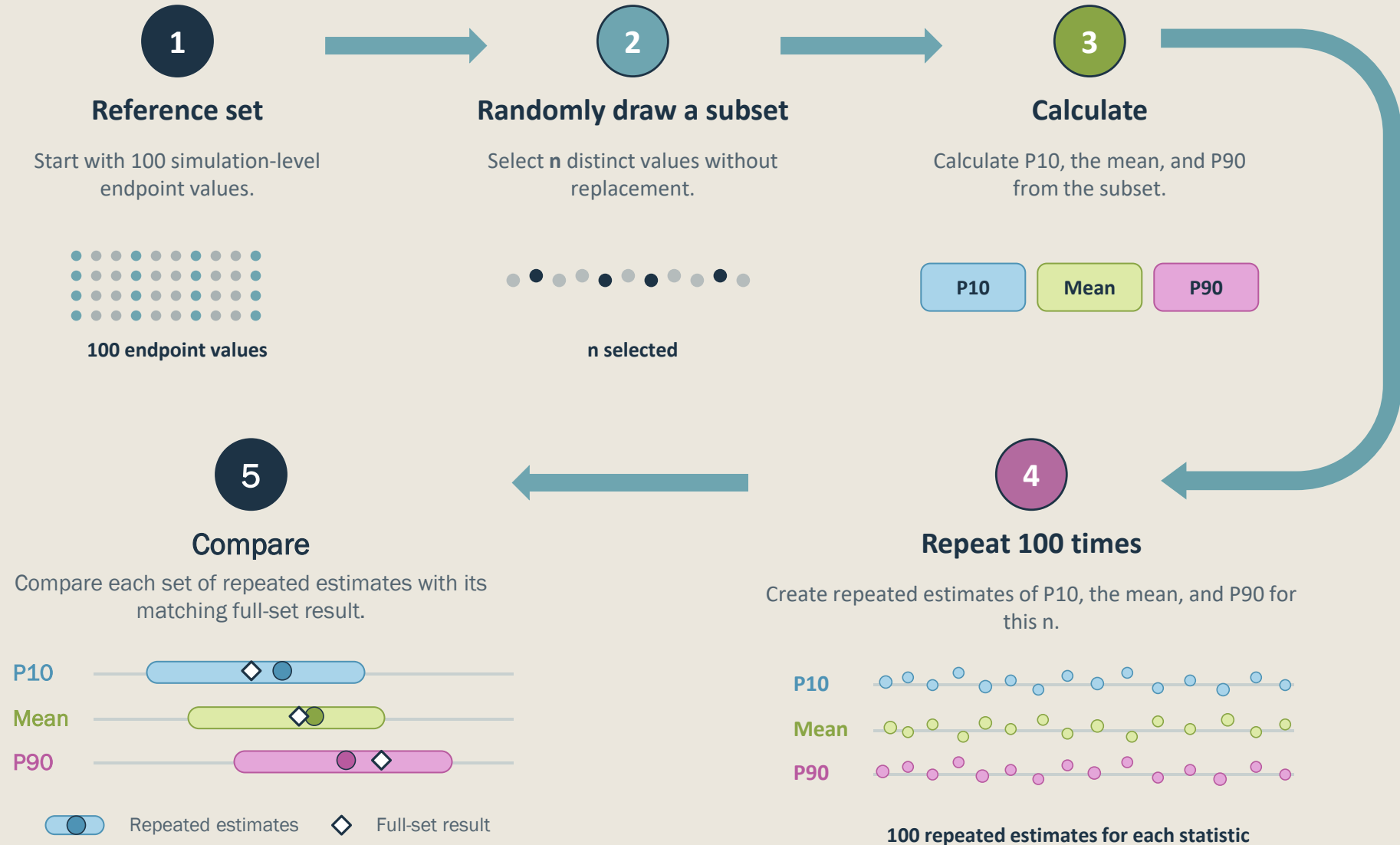
- The following analysis was conducted to determine the impact of number of simulations on assessment-related endpoints.

One “percent <1.0 mg/L” endpoint per simulation

- We applied 1.0 mg/L uniformly as a fixed reference point for this illustrative Deep Channel segment to create a consistent simulation-level endpoint for the stability analysis—not to characterize water-quality attainment.
- Across 100 simulations, the endpoint values form the distribution shown in both panels.
- P10, the mean, and P90 summarize that distribution.
 - P10 | Mean | P90: 4.662 | 4.771 | 4.924



The experiment asks how closely smaller subsets reproduce the 100-simulation summaries

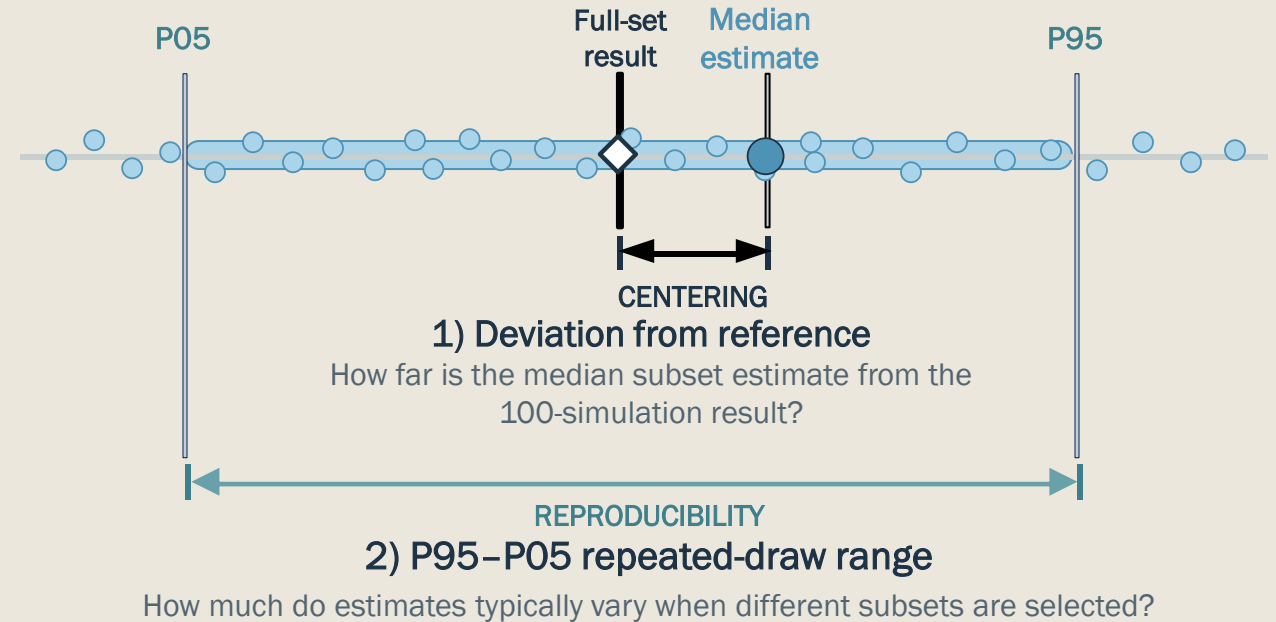


Stability has two distinct parts

- **Deviation from reference:** Is the median repeated-subset estimate centered on the 100-simulation result?
- **P95–P05 repeated-draw range:** How much do estimates typically vary when different subsets are selected?
- P10 and P90 describe endpoint values within an ensemble; P05 and P95 describe estimates across repeated subset selections.
- An estimate can be centered but still sensitive to which simulations are included.

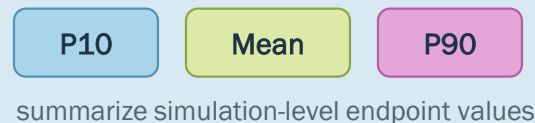
For each statistic—P10, mean, or P90—the 100 repeated subsets produce a distribution of estimates.

Consider the repeated subset estimates for **one** statistic

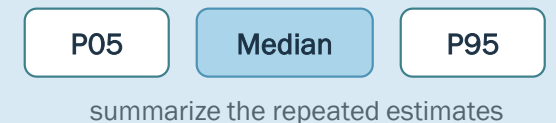


The percentile labels refer to different levels of the experiment

Within each selected subset

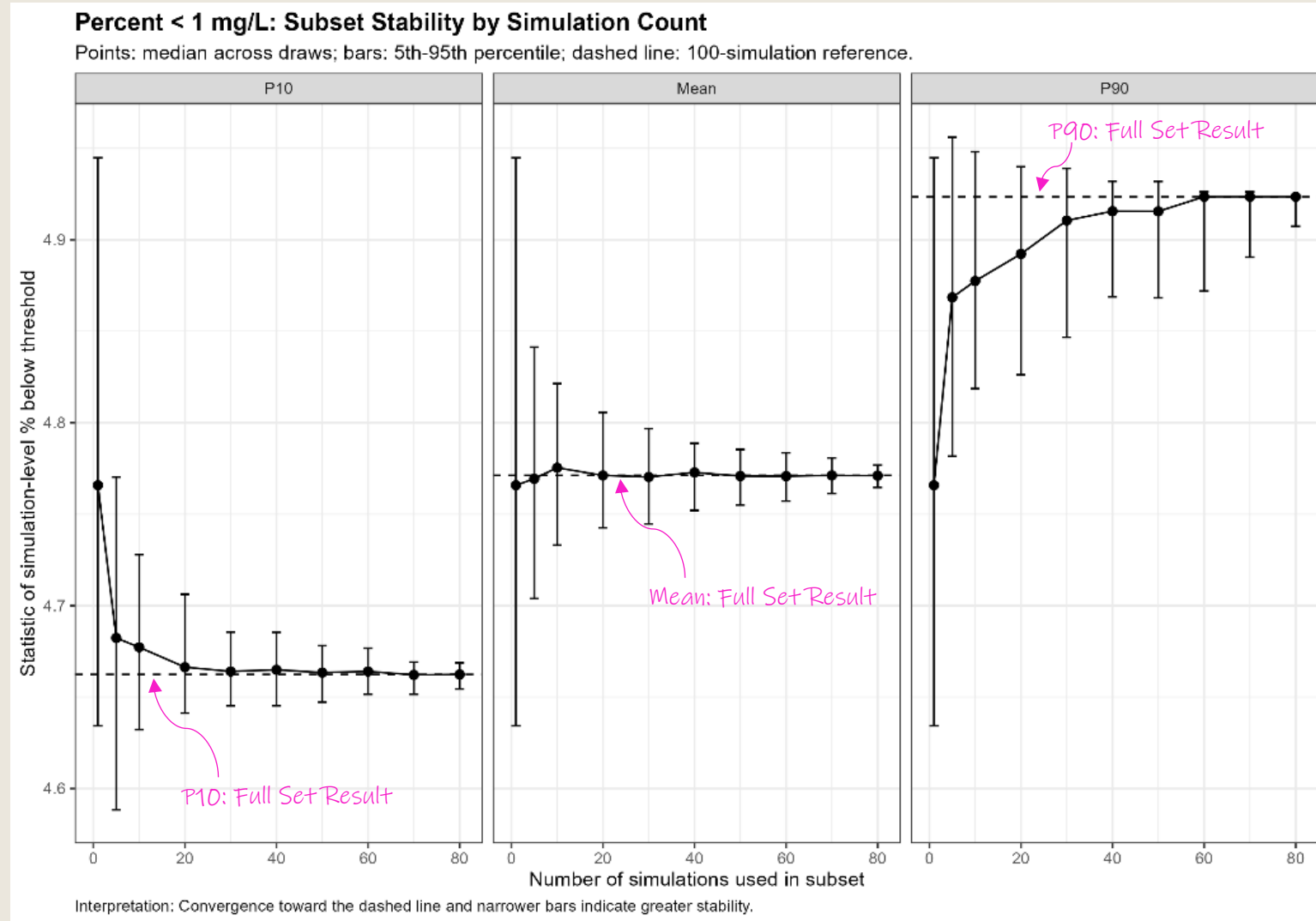


Across the 100 repeated subsets



Larger ensembles improve both parts of stability

- Median subset estimates generally approach the 100-simulation reference as n increases.
- Repeated-draw intervals generally narrow because different subsets produce more similar estimates.
- The mean generally stabilizes faster than P10 and P90.
- The tail summaries therefore have greater influence on the simulation-count decision.



The analysis uses three non-overlapping classes based on deepest designated use

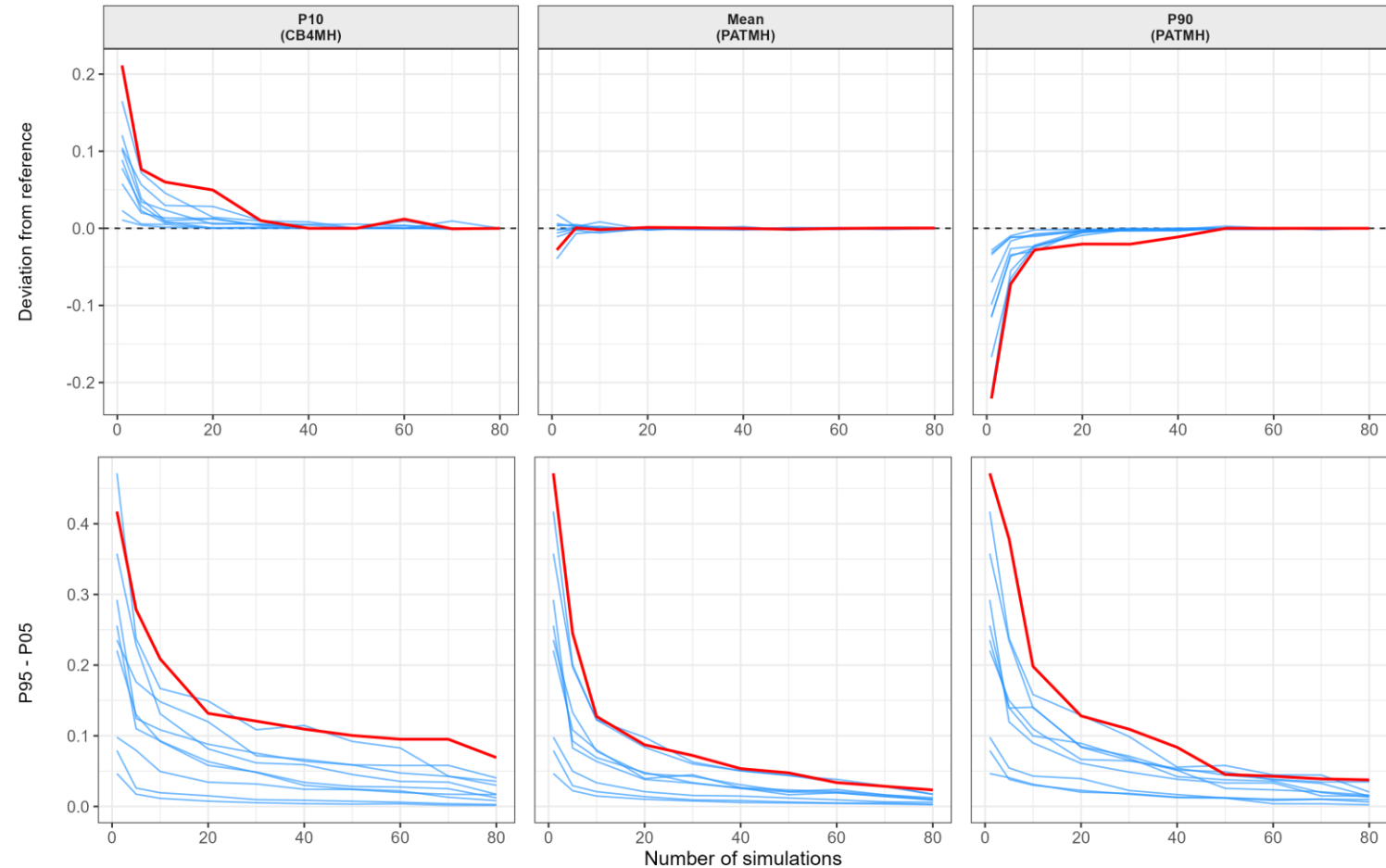
- Segments were assigned to mutually exclusive classes so that each segment contributes to only one class summary.
- The analysis includes 10 Deep Channel segments, 8 additional Deep Water segments, and 74 remaining Open Water segments.
 - *The stability summaries retain 72 Open Water segments after excluding APPTF and NORTF because their below-threshold responses were trivial.*
- Fixed reference thresholds of 1.0, 1.7, and 3.2 mg/L were used to create consistent diagnostic endpoints for the three classes.
- These endpoints were used to evaluate simulation-count stability, not to reproduce the complete criteria-assessment procedure or determine attainment.

Deep Channel class stability

- Each line represents one Deep Channel class segment; columns show P10, the mean, and P90.
- The upper row shows signed deviation from each segment's 100-simulation reference.
- The lower row shows sensitivity to subset selection, measured by the P95–P05 repeated-draw range.
- Centering improves relatively quickly, while reproducibility improves more gradually.
- At 50 simulations, the median repeated-draw ranges are 0.037, 0.021, and 0.035 percentage points for P10, the mean, and P90.

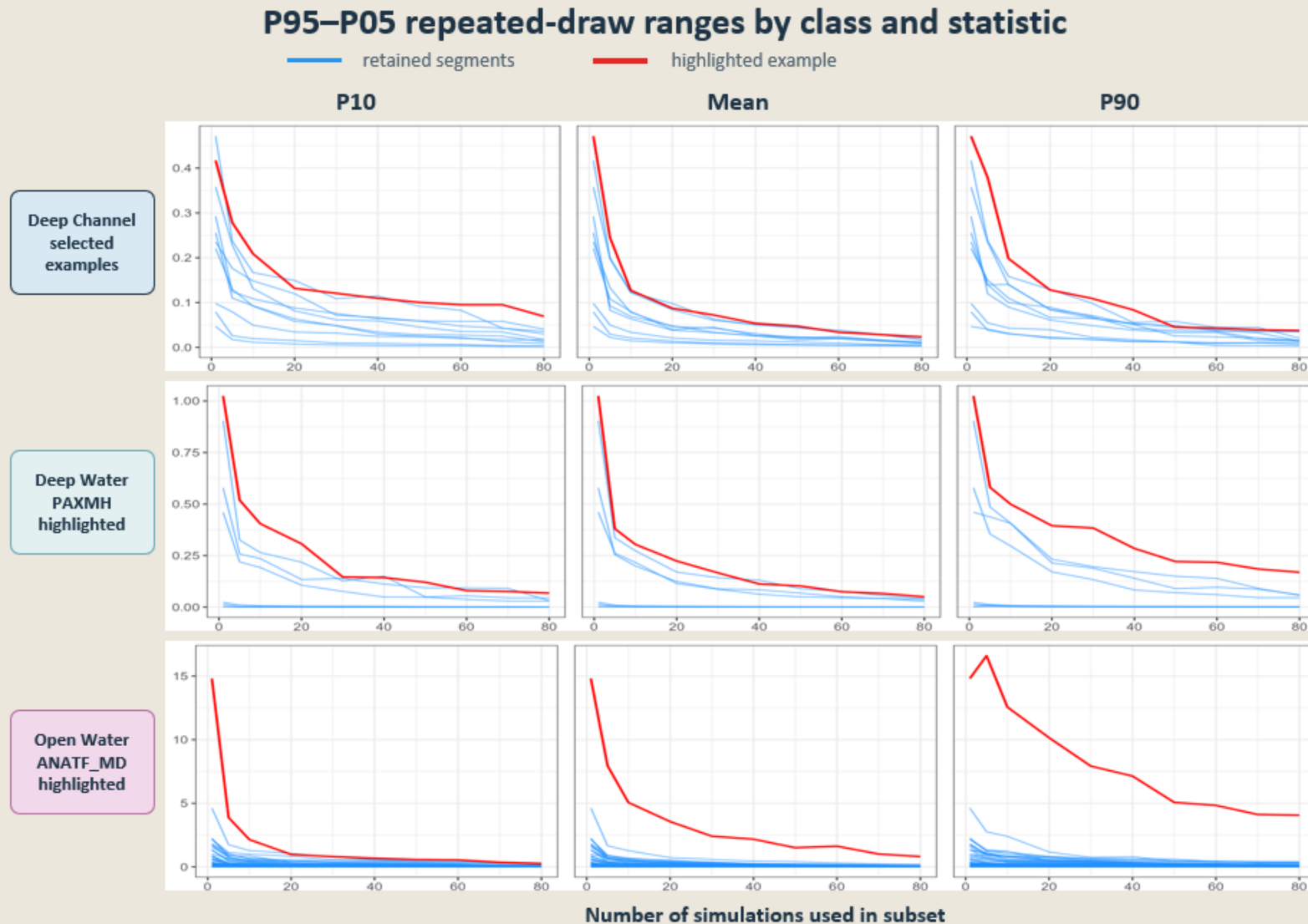
Subset Stability Summary by Class (Deep Channel - Approach C retained set: <1.0 mg/L)

Red line shows the highlighted segment named in each facet strip.



20 simulations support testing, 50 simulations improve stability

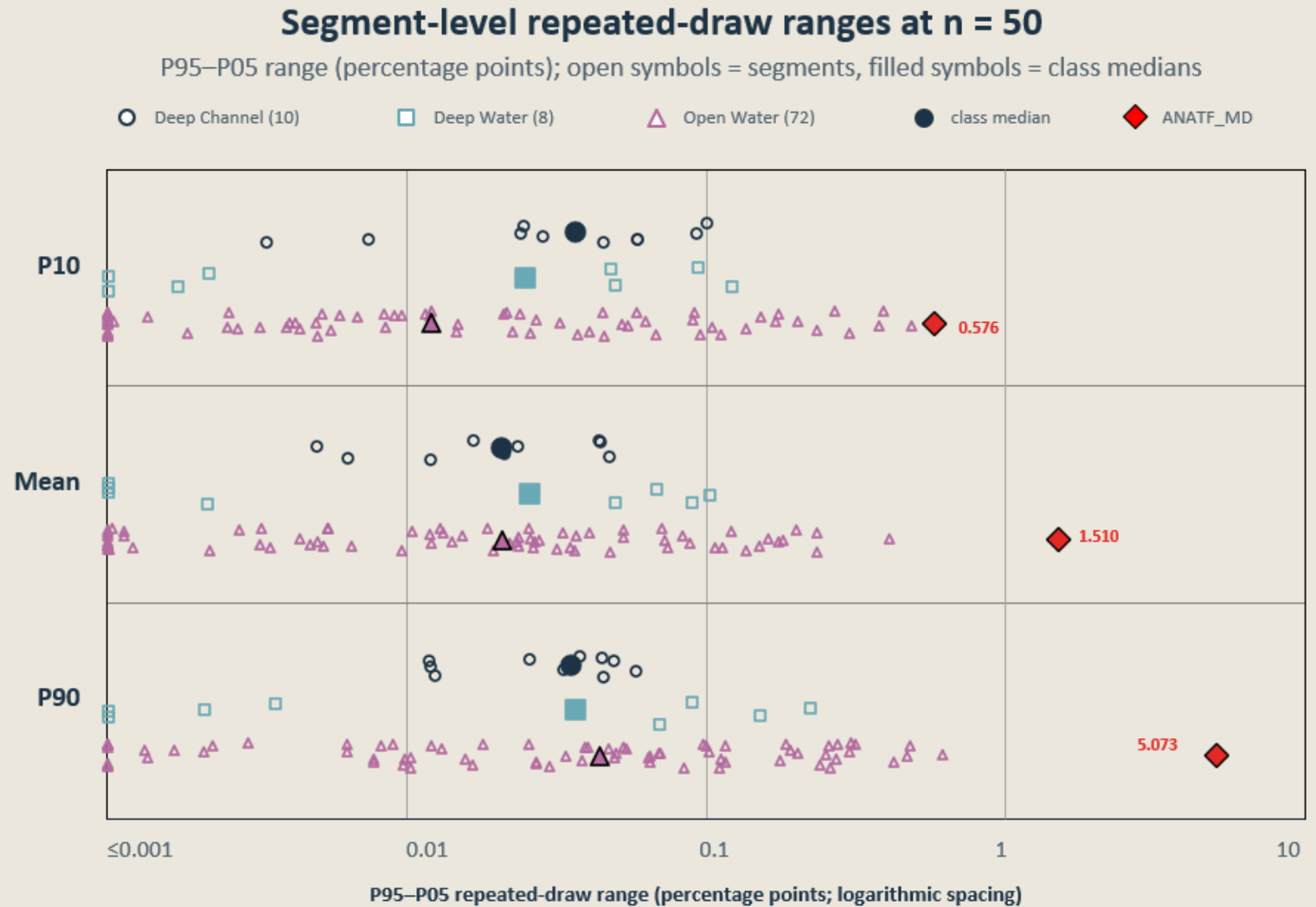
- By 20 simulations, class-median **absolute deviations** from the reference were below 0.01 percentage point.
- Across classes and statistics, class-median repeated-draw ranges narrowed from **0.036–0.088 percentage points at n=20** to **0.012–0.044 at n=50**.
- P95–P05 repeated-draw ranges generally continued to narrow, but class medians **changed by no more than 0.006 percentage point** from 50 to 60 simulations.
- ANATF_MD is an exception to the general pattern; later slides examine its sensitivity in more detail.



Ranges are percentage points; each class row uses its own vertical scale.

Most segments are stable at 50 simulations

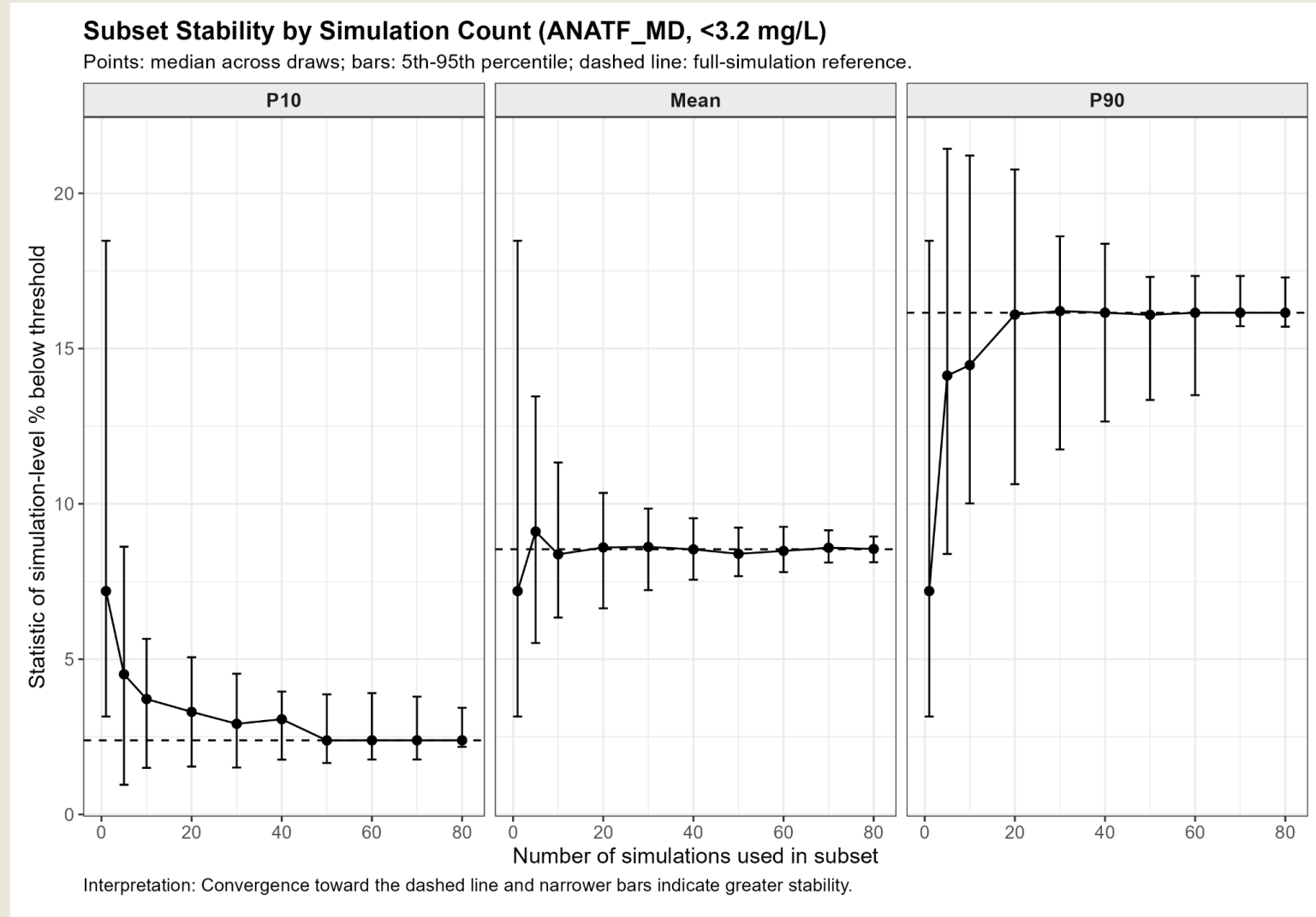
- **Open symbols**—individual segment repeated draw range for n=50.
- **Filled symbols**—median repeated draw range.
 - P10: 0.012–0.037
 - Mean: 0.021–0.026
 - P90: 0.035–0.044
- ANATF_MD is the clear exception, with ranges of 0.576, 1.510, and 5.073 percentage points for P10, the mean, and P90.



Values ≤0.001 percentage point—including four zero P10 ranges—are shown at the left bound.

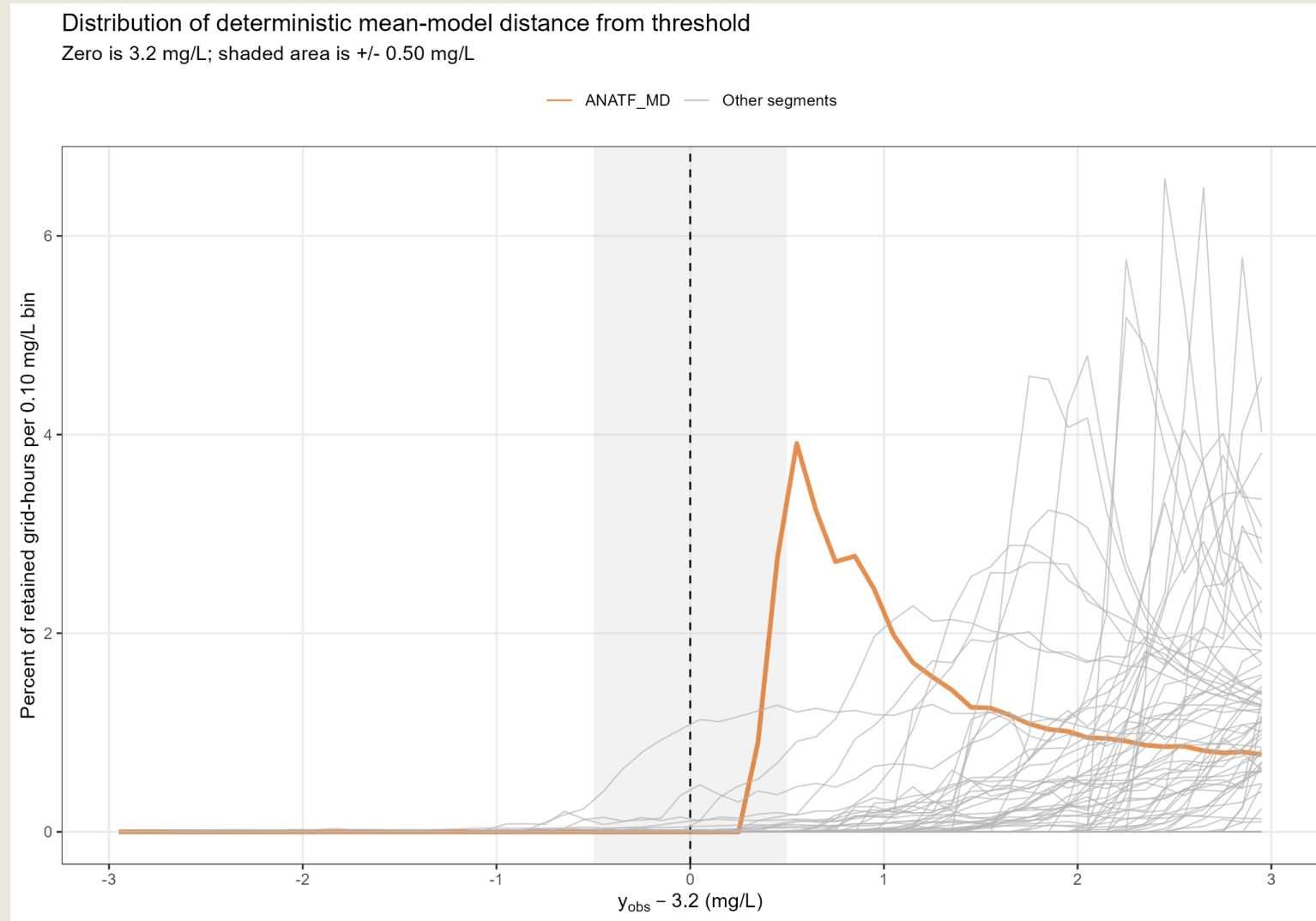
ANATF_MD is a stress case, not the basis for the standard count

- The median repeated-subset estimates approach the 100-simulation reference.
- Repeated-draw intervals remain wide, especially for P90.
- At 50 simulations, the ANATF_MD P90 range is 5.073 percentage points, compared with an Open Water class median of 0.044.
- ANATF_MD therefore requires separate scrutiny, but it should not determine the standard ensemble size for all segments.



Threshold proximity helps explain the ANATF_MD result

- ANATF_MD has an unusually large share of deterministic predictions near the 3.2 mg/L reference threshold.
- Approximately 3.7% lie within ± 0.50 mg/L of the threshold—second highest among the 74 Open Water segments.
- Approximately 18.7% lie within ± 1.00 mg/L—highest among the 74 segments.
- Comparatively limited grid-hour support may also contribute; residual variability is moderate rather than uniquely extreme.
- The sensitivity appears to reflect several contributing conditions rather than one isolated cause.



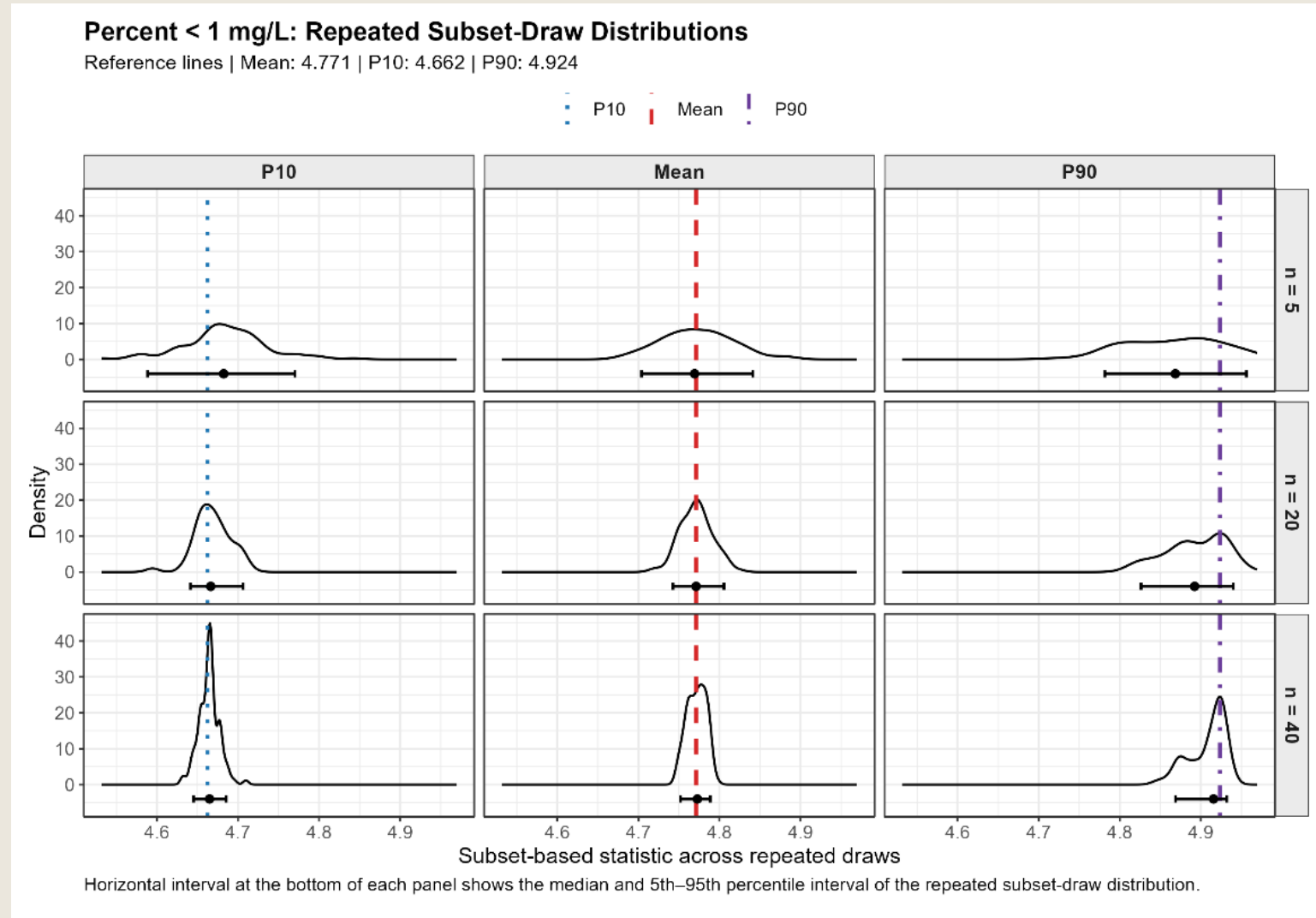
Synthesis and application

- The mean stabilized sooner than P10 and P90, so the tail statistics have greater influence on the production count.
- At 20 simulations, the typical segment's estimate was already close to its 100-simulation result, but tail estimates and sensitive segments were more variable.
- Use **21 simulations for workflow testing and exploratory work**, not for production analyses.
- For most segments, improvements became progressively smaller around 50–60 simulations; use **51 as the provisional production count**.
- ANATF_MD results might suggest increased simulation size for borderline assessments.
- Re-evaluate the production count using the adopted CFD procedure and percentile-based assessment rule.

Extras

Tail stability using an Illustrative Deep Channel segment

- The mean narrows relatively quickly; P10 and P90 remain more sensitive to which simulations were selected.
- Increasing n from 5 to 20 to 40 improves both centering and reproducibility.

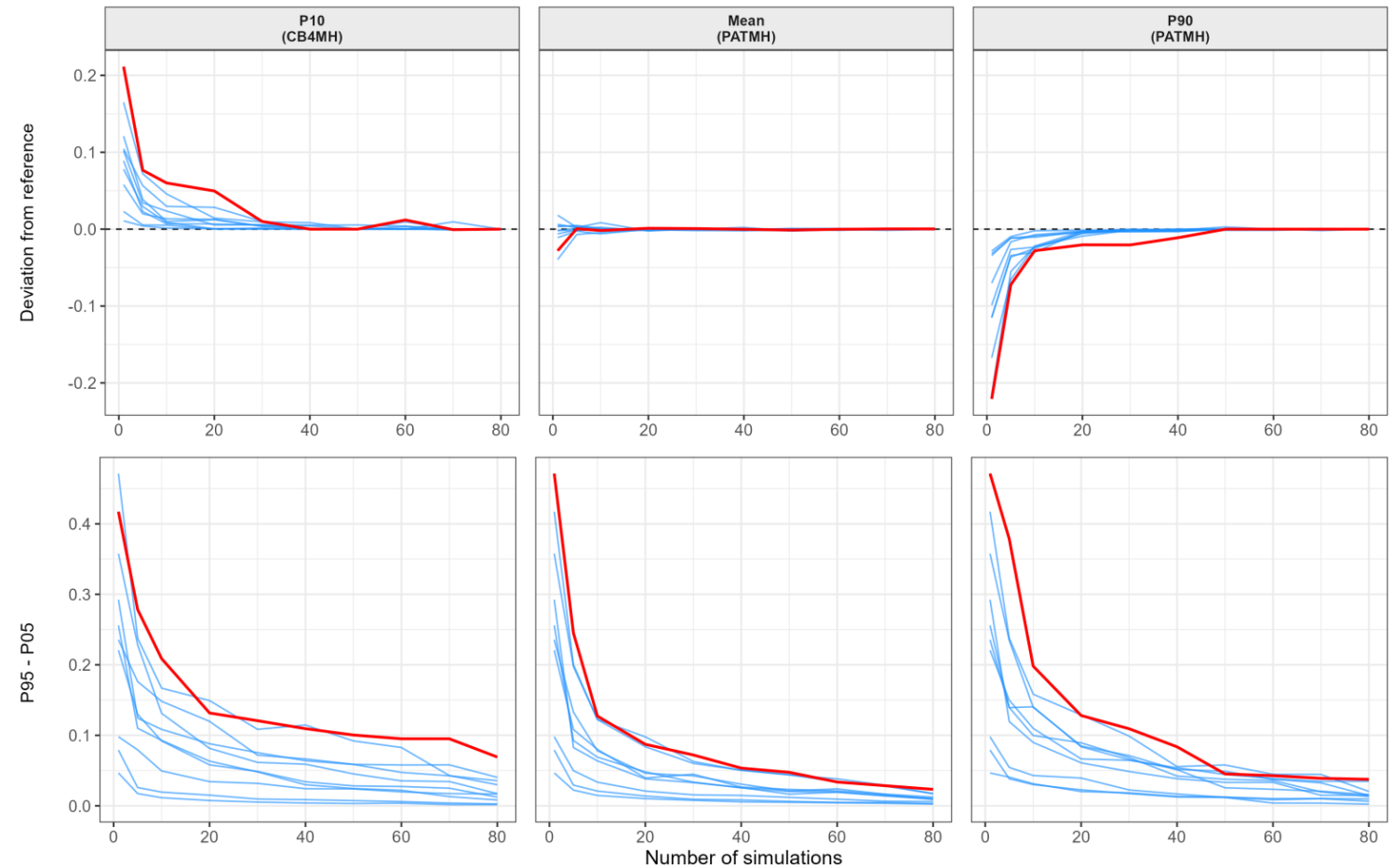


Deep Channel Class

- At 50 simulations, median P95–P05 ranges are 0.037, 0.021, and 0.035 percentage points for P10, mean, and P90.
- The additional change from 50 to 60 simulations is modest.

Subset Stability Summary by Class (Deep Channel - Approach C retained set: <1.0 mg/L)

Red line shows the highlighted segment named in each facet strip.

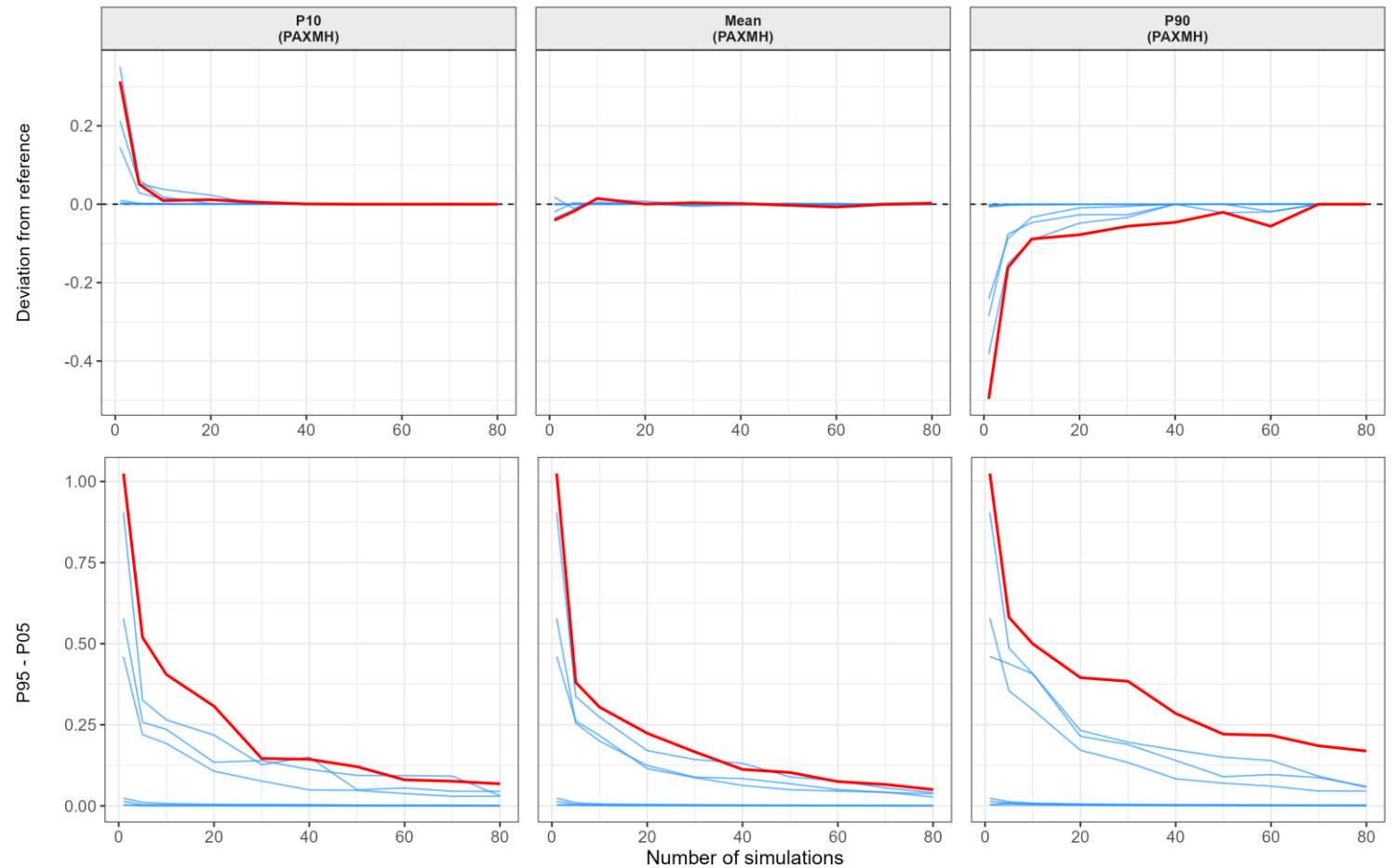


Deep Water Class

- At 50 simulations, median ranges are 0.025 (P10), 0.026 (mean), and 0.037 (P90) percentage points.
- PAXMH's P90 repeated-draw range remains comparatively high: 0.221 at n=50 and 0.217 at n=60.

Subset Stability Summary by Class (Deep Water - Approach C retained set: <1.7 mg/L)

Red line shows the highlighted segment named in each facet strip.



Open Water Class

- At 50 simulations, median P95–P05 ranges are 0.012, 0.021, and 0.044 percentage points.
- ANATF_MD's repeated-draw ranges reach 0.576, 1.510, and 5.073 percentage points for P10, the mean, and P90.

Subset Stability Summary by Class (Open Water - Approach C retained set: <3.2 mg/L)

Red line shows the highlighted segment named in each facet strip.

