The Department of Ecology's Supplemental Modeling
Report. A Critical Review.
David H. Milne PhD July 2018

3. THE COMPUTER GETS MANY WRONG ANSWERS.

3-1. Overview.

Appendix G2 of the original TMDL Report presents 38 pages comparing the Budd Inlet Model's output with the observed water quality parameters that were used to calibrate it (TMDL Appendix, 2012). There are three pages for each of the Appendix G2 stations, each portraying observed and calculated conditions at the surface, bottom, and a depth midway between surface and bottom. Figure 3-1 shows an example, this one for the dissolved oxygen levels in the bottom water at station BI-6 in West Bay (the station nearest the dam).

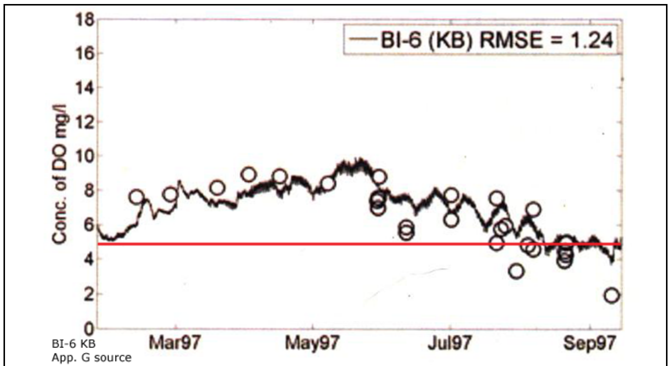

Figure 3-1. Observed dissolved oxygen levels (circles) and Budd Inlet model's calculations (graph line) of those levels near the head of West Bay (station BI-6), in water at the bottom. January 25 – September 15 1997. Red ine shows the DO standard at that station (5.0 mg/L) Source: Appendix G2 TMDL Report, page 1.

These Figures show a remarkable ability of the model to follow (and roughly predict) the observed levels of dissolved oxygen in the water over the simulated "year" (January 25 - September 15, 1997). In those Figures, the computer's graph (dark line) follows the general trend of the observed data (open circles) quite faithfully. However if *every* calculation were accurate, the graph would go through and touch the center of *every one* of the open circles. It doesn't do that. It "misses the mark" by a wide margin in some cases, by a narrow margin in others, and in some cases (where it passes very close to the centers of the data circles as it goes through them) it is "dead-on accurate."

The differences between the positions of the data points and how far the graph is above or below each one is a measure of the average size of the error made by the computer. As Figure 3-1 shows, the computer's graph passes directly through *almost none* of the observed data points. The errors are large. How large is explored in the Sections that follow.

That is the fact to always bear in mind; the computer often gets wrong answers. *Yet the Ecology modelers interpret its outputs as though every one of the thousands of calculations is dead-on accurate.*

3-2. Counting Right Answers.

The data points in the Appendix graphs are at the exact centers of the circles shown there. These circles are about 0.875 mg/L in diameter. If the graph fails to touch ("misses") the circle, the computer's answer in that case is in error by at least 0.44 mg/L (the circle's radius). That is about twice the critical threshold (0.2 mg/L below the water quality standard) used by the modelers in judging whether a violation has been detected.

I examined each of the dissolved oxygen graphs in Appendix G2 (36 graphs; 3 depths for each of 12 stations) for visual determination of whether the computer graph missed the observed data point circle, "hit" it, or was undeterminable (not clearly a hit or miss). To qualify as a "hit," the graph had to touch the exact top or bottom of the data circle or pass through it. A grazing contact along one side of a circle was scored as a "miss;" the graph was close in that case but the top or bottom (on a vertical line through the center) of the data circle was not in contact with the graph on the date of the observation. Figure 3-2 for station BF-3 surface water (near Boston Harbor) shows an example. Figure 3-3 summarizes the "hit" and "miss" pattern for all 36 graphs.

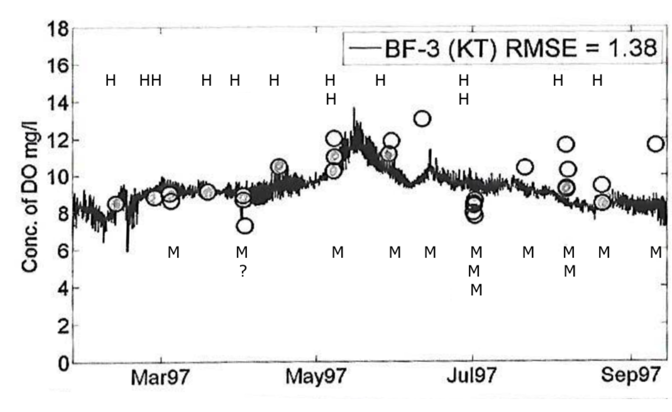

Figure 3-2. Assessment of calculated "hits" and "misses" of observed data circles by the Budd Inlet Model for dissolved oxgyen concentrations in surface water at station BF-3 (near Boston Harbor) by the method described in the text. Hits ("H" in upper row), misses and undeterminables ("M" and "?" in lower row) show 13 accurate, 13 inaccurate and 1 undeterminable calculation. Source Appendix G2 p. 36 TMDL Report.

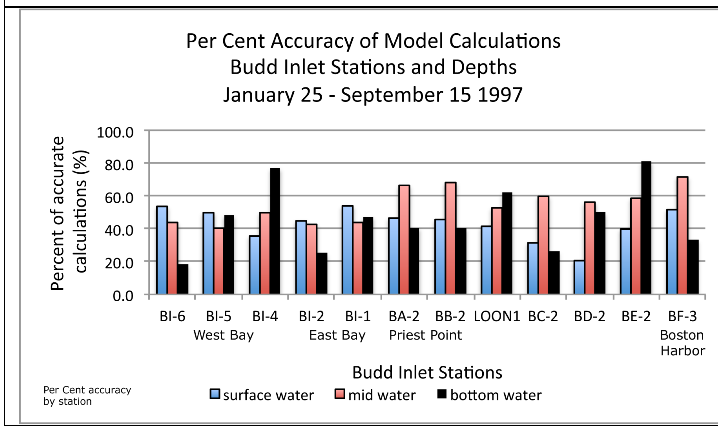

Figure 3-3 shows that the computer's calculations matched observed DO's about 80% of the time in bottom water at sites BI-4 (entrance to West Bay) and BE-2 (center Budd Inlet near the Tamoshan area). At worst, calculations matched the observed values in bottom waters only about 20% of the time at BI-6 and BI-2 (West and East Bays) and BC-2 (Gull Harbor area). Overall, the calculations were accurate in roughly 40-50% of cases (Fig. 3-3).

If the Budd Inlet Model calculates wrong answers (estimates of DO levels) that are always very close to the "right answers" (actual real-life DO levels), that is still very helpful and informative. But it doesn't consistently do that. When it "misses the mark," it can do so by a large margin of error. The following shows one calculation error that could hardly be worse, and that in general the calculations can be far from accurate.

Figure 3-3. Accuracy of the Budd Inlet model. Bars show the per cent of calculations that correctly identified observed DO values (counting all "indeterminable" scores as "hits") by stations from south to north in Budd Inlet. Data from graphs in Appendix G2 TMDL Report.

3-2a. Worst Case Scenario.

In the preceding, I used only data from the modelers' own graphs in Appendix G2. The inability of the model to "get it right" in every calculation is also evident if data from other sources are used. Figure 3-4 (same as Figure 3-1 above) shows the bottom water at station BI-6 with an added overlay of data points from the BISS spreadsheet for that site. The data presented by the modelers (circles) are identical to those from the spreadsheet (triangles) in many instances. The modelers' data include values not found by me in the spreadsheet (for example, two points near July 1 whereas the spreadsheet shows only one) and values found in the spreadsheet that are not shown on the modelers' graph (for example, the very high data point in mid-September).

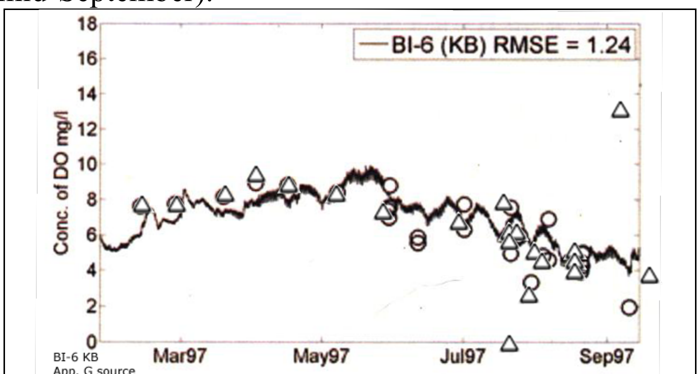

Figure 3-4. Figure from the TMDL Appendix with an overlay of data from the BISS research (triangles). The rightmost triangle is September 24, a few days after the end of the computer simulation interval. Source: BISS spreadsheet.

The lowest observed DO level shown in Figure 3-4 is ~2.0 mg/L (circle at September 10). It was evidently part of the data set used to calibrate the model. The BISS spreadsheet shows no such number but instead lists the bottom water DO level for that date as 12.53 mg/L. That high level of DO was actually observed, as were similar high bottom water DO's at two East Bay sites on the same date. This situation (explained in detail in Chapter 5) was due to intense late-summer photosynthesis by microscopic algae attached to the shallow sunlit bottom.

An accurate simulation of the benthic algae's DO production would have enabled the computer to "see" it. (The line traced by the computer would have "shot up" to 12.53 mg/L on that date, then back down again by the next day.) It did not. In fact, the graph drops to its lowest level of the entire season on that date (Figs. 3-4 and 3-1). T*he computer model predicted the lowest bottom water dissolved oxygen of the entire season on a day when the benthic dissolved oxygen was actually at its highest level of that season*.

The same giant error occurs at station BI-1 in East Bay (see Figure 2-2 for location). The patch of water immediately to the east of BI-1 (actually touching the BI-1 grid square) is Ecology's "critical cell" – the grid square that almost always shows the worst (lowest) seasonal dissolved oxygen levels of the entire year. Here is the situation that the agency blames on Capitol Lake. And exactly here is where a component of the computer model (the "benthic algae subroutine") failed catastrophically, showing a low bottom water DO level when it should have showed the highest level of the entire year.

3-2b. All Model Calculations for the Critical Cell; Flawed?

This is no small laughable error. It implies that *all* of Ecology's DO calculations for East Bay are suspect, as they may well be for all other shallow areas around Budd Inlet where benthic algae are at work, all summer long, wherever else the benthic algae subroutine failed.

A low DO "observed" value is shown in Fig. 3-1 (West Bay) on September 10. That low value is listed in the BISS spreadsheet for that time and place, but it is identified as an error in that spreadsheet. No low "observed" September 10 values are shown in the graphs for East Bay, however those graphs are also lowest on that date.

What about observed data other than the BISS record for East Bay? To my knowledge, there aren't any. Despite its critical central role in the Budd Inlet modeling effort, the Department of Ecology has never (to my knowledge) actually measured DO levels at that location. Nor have LOTT staff, nor has anyone else – to my knowledge. The 1997 data collected during the BISS study are the last ones ever made there. *All* of Ecology's pronouncements on the alleged negative effect of Capitol Lake on that location are based on demonstrably flawed computer calculations.[1]

3-3. The Computer's Margin of Error.

Every graph in Appendix G2 (for example, Fig. 3-4) shows a number at its upper right hand corner labeled "RMSE." This is the "Root Mean Square Error," the computer's "margin of error" for that location and depth.[2] These numbers are quite large, ranging from 0.52 mg DO/L (bottom, site BE-2) to 4.72 mg/L (surface, BB-2).

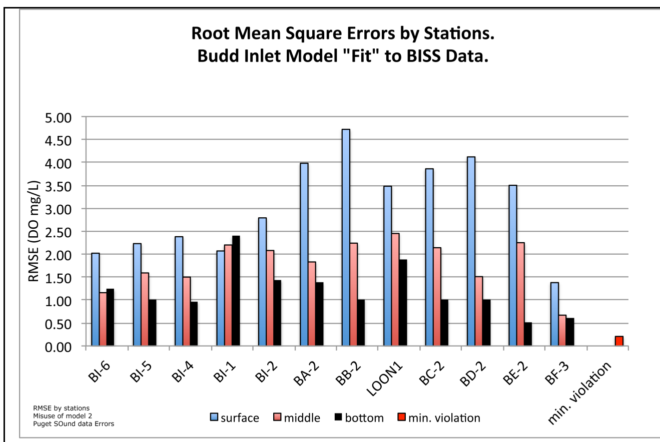

Figure 3-5. Values of the model's "margin of error" (RMSE) in surface, middle depth, and bottom waters at Budd Inlet stations. West Bay is to the left, Boston Harbor is at right. Far right; size of the smallest DO below DO standards that qualifies as a violation. Source: TMDL Appendix.

The RMSE for each situation is calculated by averaging the differences between each known observed DO level (circles) and the computer's estimates of those observed values

---

[1] That said, I expect that if measurements were actually made there, they would show low DO levels at the bottom. East Bay's backwater situation is a textbook example for occurrences of low DO's, for reasons described in Chapter 1.

[2] The RMSE is defined and briefly discussed on p. 57, TMDL Report. I asked the modelers whether "margin of error" is an accurate interpretation of the RMSE during our November 2014 meeting. They said "yes."

(graph lines). The RMSE is thus the average difference between the real values and the calculated estimates of those values. Figure 3-5 shows the sizes of the RMSE's at all sites and depths, and the (small) size of the minimum violation (0.2 mg/L) that the modelers seek to detect using a tool with those large margins of error.

Figure 3-6 shows the likelihood of making a mistaken decision about DO standards violations in four possible situations ("scenarios").

For station BI-6 (closest to the dam; bottom water RSME = 1.24 mg/L [Fig. 3-1]) the DO standard is 5.0 mg/L. Violations occur if the DO falls 0.20 mg/L lower than this, that is lower than 4.8 mg/L.

For a real-life DO of 6.24 mg/L (scenario 1, Fig. 3-6), the average "low side" calculation error by the computer would be one RMSE (= 1.24) lower than this, namely 5.0 mg/L. That is higher than the 4.8 mg/L "cutoff" or threshold for declaring a DO standards violation. When the real-life (but unknown to us) DO level is that high, the computer will almost always correctly recognize "no violation."

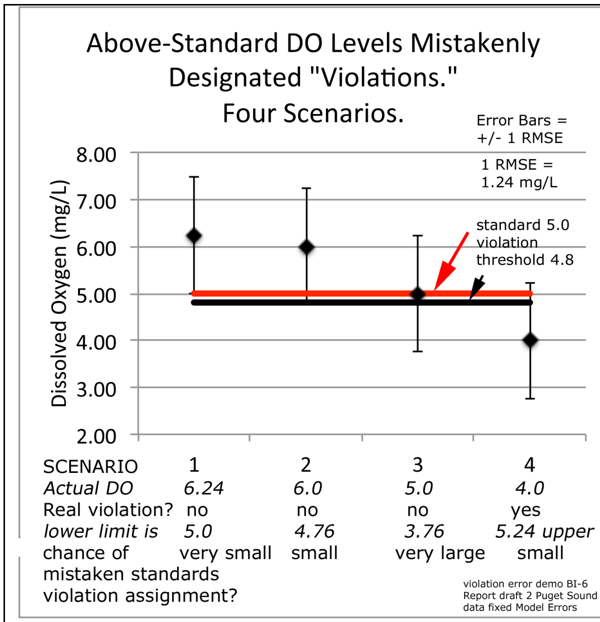

Figure 3-6. Diagram illustrating how unknown DO levels close to the DO standard can be mistakenly designated as "violations" or "no violations" due to the computer's large margin of error.

As the real-life DO level declines, however (scenarios 2, 3, and 4), the "slop" in the computer's predictions overlaps the violation threshold by more and more, increasing the likelihood of "finding" violations when there are really none. Finally, for DO's below the threshold (scenario 4), "high side errors" by the computer could result in the mistaken assignment of "no violation" to a situation in which a violation really has occurred. The closer the real-life DO is to the DO standard level, the more likely it is that the model's calculation will be in error one way or the other – "violation" when there is none, or "no violation" when there really is one.

Until recently,[3] the "violations" maps used by Ecology to show the computer model's output were like the leftmost diagram in Figure 3-7. Each colored square is a location where the computer "found" at least one DO level lower than the DO standard assigned to that location during its exhaustive search between simulated January 25 and September 15. If more than one violation was calculated for a site, the color shows the size of the

[3] The Dept. of Ecology may have switched to another form of map during 2017. See Figure 3-9, this Chapter.

most serious violation – the “worst case” of the “year” at that place. The scale alongside the map shows the sizes of the violations represented by the colors.

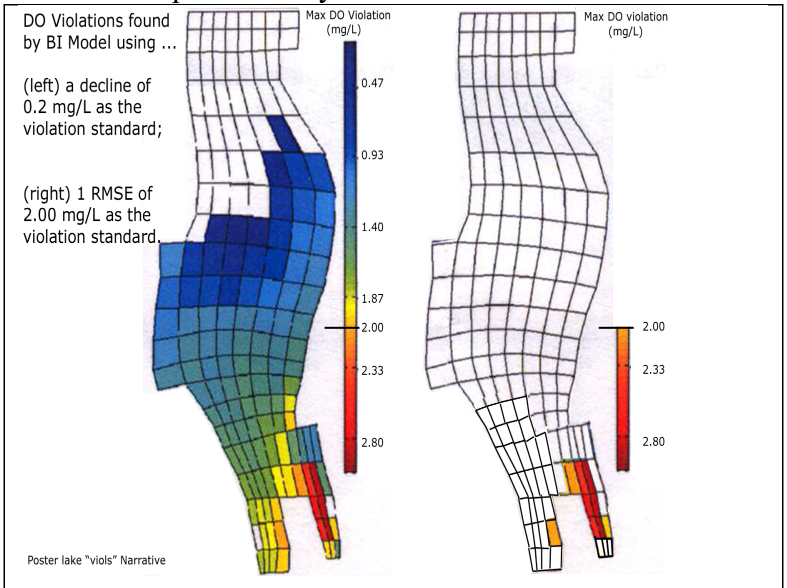

Figure 3-7. Areas in Budd Inlet flagged for violations where calculated values are (a, left) 0.2 mg/L or more below the standard or (b, right) one RMSE (in this example, 2.0 mg/L) below the standard. The modelers’ colored scale (a) shows the maximum amount by which each colored square is in “violation.” Source: a) Poster 2014; b) Author’s calculation presented in Power Point OK2, 2014.

The uppermost end of the scale, not numbered by the modelers, is at 0.2 mg/L, the smallest possible violation.[4]

The darkest blue violations on the map are all 0.2 mg/L. These microscopic “violation” were the worst that the model was able to calculate at those sites. Its margins of error for sites in this central region are all much larger than 0.2 mg/L (Figure 3-5). In real life (as discussed in Chapter 8) the measured DO’s at those sites during the BISS research were never lower than the standard for that region (6.0 mg/L) during the months simulated. It is certain that many of the blue “violations” shown resulted from large random low-end errors of estimates in a few of the many thousands of calculations made for those sites – but the modelers regarded every last one of those calculations as accurate.

The rightmost map in Figure 3-7 shows a way of screening out some of the uncertainty created by the model’s large margins of error. In that diagram, only calculated violations that are lower than the DO standard by one RMSE or more are shown. (For this illustration I used 2.0 mg/L as the RMSE for the whole inlet.) A departure that large seems likely to indicate a real violation rather than a low-end error by the computer. The resulting violations map shows far fewer colored squares – but we can be more confident that the ones shown are not just portrayals of random errors made by the computer.

3-4. The Biggest Source of Error and Confusion of All.

Figure 3-8 shows the oxygen standards for Budd Inlet water quality. They are 5.0 mg/L in the southern harbor sector and 6.0 mg/L over the larger central and northern sectors. A violation occurs if the real life DO level drops below these standards by 0.20 mg/L or more (that is, below 4.8 or 5.8 mg/L).

[4] … found by me by using a ruler and extrapolation ...

There is a loophole, however, in this straightforward comparison. If the "natural" water as it existed before intensive human activities began was already below the modern standards, then the level of oxygen in that pre-modern water itself becomes the standard. Thus for pre-modern water that occasionally had a DO level of 3.8 mg/L in its natural state, for the part of the season when that low DO occurred the DO level of the modern water would have to drop 0.2 mg/L below 3.8 (that is, to 3.6 or lower) before a violation is declared. This exception protects modern waters that once ran low on DO due to natural causes from attracting undue regulatory wrath for conditions that human activities did not create. It is clearly described on page 35 of the SPSDOS (2013) Report.

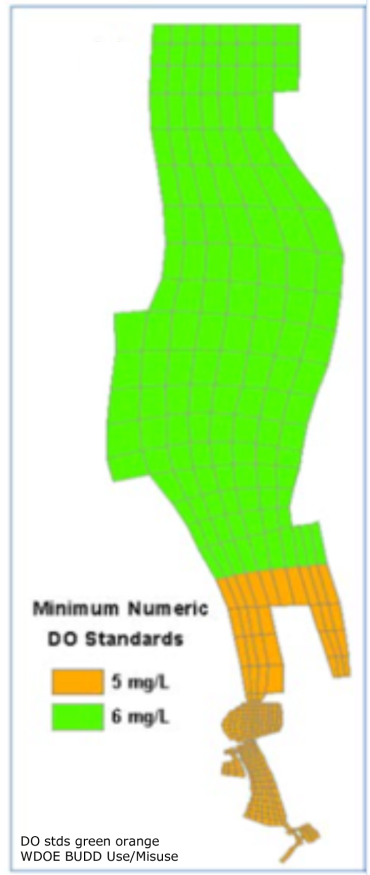

Figure 3-8. The water quality standards for dissolved oxygen in Budd Inlet; 6.0 mg/L in the green sector, 5.0 mg/L in the orange area. The 5.0 standard is used for the Capitol Lake basin in simulations of the Inlet in its "natural" configuration before the dam was built. Source: SM Report Fig. 7, p. 32.

A practical problem with this exemption is that very few measurements of natural pre-modern waters were made before human activity became intense. For example, I know of no DO measurements in Budd Inlet earlier than 1957, by which time the Deschutes estuary had already been dammed. We seldom have data from the pre-modern era, in which case we must simply use the modern numerical standards of Figure 3-8 as our guidelines. At the time when the SM Report was printed (2015) the modelers were using this "natural waters" loophole on a grand scale. They asserted that they could "know" what the pre-modern Deschutes estuary was like by using the Budd Inlet Model with inferred data from the past – estimated stream flows, estimated nitrate concentrations, estimated climate and weather, and the like.[5] The model output grid map shown in Figure 3-7 above actually resulted from two parallel runs – a simulation of the "natural" Budd Inlet estuary and one of modern Budd Inlet -- compared point by point every step of the way. The "violation" shown in each colored square could be from when the natural inlet had low DO's and the modern inlet dropped even lower, or where the "natural" inlet water was above the modern standards but modern waters dropped below the standards 6.0 or 5.0 mg/L. The map gives no hint of which situation resulted in the calculated "violation" shown in each square.

[5] The modelers refer (SM Report p. 26) to TMDL Appendix I for 'natural' conditions of the past. Confusingly, Appendix I (p. I-7) says that "current" values of the Deschutes River flow – and temperatures and other properties – were used in their simulations of 'natural' pre-modern waters. This is in stark contrast to their reply to my questions about this (see Chapter 9, this Review).

This is the most error-prone procedure of all. The model with all of its leeway for error is first used to decide what the natural estuary was like, then it is used again to compare the modern Budd Inlet with the natural estuary. Probabilities of success get squared in such procedures; for example if the chance of "getting it right" once is only ½ , the probability of getting it right twice is only ½ x ½ = ¼.

This procedure also disguises the severity of the oxygen shortages caused by the calculated "violations." In the dark blue center of Budd Inlet (Fig. 3-7), is the "worst case" DO level at 5.8 mg/L (standard 6.0 less 0.20)? Or is it at, say, 3.6 mg/L ("natural" low DO, say 3.8, less 0.20)? The former would still be "good" water quality, the latter would be critically low in oxygen.

In summary, the "natural water exception" magnifies the probability that the computer produces wrong answers. *It also prevents outsiders from checking the answers*. To do so they would need a copy of the model and a computer and staff comparable to Ecology's to compare the grid maps for the modern and natural estuaries – resources not available to the public and other interested parties.

<u>3-5. Minimizing Errors; Possibilities and Ecology's Responses.</u>

Is there a way to be more confident that the "violation grid maps" really identify locations that are likely to have DO violations? Yes. The answer is to trust averages of numbers, rather than every individual number by itself. For example, suppose that all of the calculated DO values for, say, location BI-1, surface water (0 m depth) and date Sept. 15 for the 6-hour interval centered on a high tide of that day were averaged. (That average would be the mean of about 65 calculations.) If that average was 0.2 mg/L or more below the DO standard for that location, we could be confident that real-life violations should be expected at that time, place, and depth.

For even more powerful confidence, a standard statistical technique for dealing with uncertainty (calculation of "confidence limits") might be applicable. However, a professional statistician's opinion is needed in this case.[6]

The modelers have in the past refused to resort to averages and have insisted that every individual number be taken at face value. In the SPSDOS 2013 Report (p. 35) they have said that averages cannot be used to "mask" the fact that a grid cell's DO dropped even momentarily below the DO standard for that area. They have taken each individual calculation at face value and assume that it is accurate enough for real-life policy decisions.

Their confidence in the dead-on accuracy of every calculation is shown in Report SPS-DOS (2013) on page 87. There they describe a location with a modern DO standard of

[6] Confidence limits can be used and are easy to calculate when the value obtained by each measurement or estimate is completely independent of the values obtained from all other measurements. In the computer's case, the value of each estimate is calculated from the sizes of previous measurements (that is, the values are "not independent"). Only a professional statistician can advise in such situations.

5.0 mg/L where the calculated DO of the 'natural' water dropped to 4.95 mg/L for all or part of just one day out of the 302 days simulated by their model.[7] Using the "natural water exception" loophole, they judge modern waters at that time and place by comparison with 4.95 mg/L -- not the standard, 5.00 mg/L – showing their confidence that their calculations are always accurate even to the second decimal place.

Personnel of the HDR engineering firm asked the modelers about accuracy in the firm's comments on the draft SPSDOS Report (2013). In their words:

"*Page 19: The DO decreases calculated by the model range from 0.2 to 0.4 mg/L in limited areas due to point sources. These are very modest changes in the DO levels in these locations. Due to these small calculated DO decreases, the following question arises: Is the model sufficiently accurate to predict these DO decreases? And more importantly, is there sufficient confidence in the DO decreases calculated by the model to mandate expensive nitrogen removal upgrades at point source treatment facilities to reduce nitrogen loadings?*"

The Department of Ecology did not respond to the HDR query (Clark, 2016).

<u>3-6. No More "Natural Estuary" Calculations?</u>

Calculating the DO violations in the "natural" pre-modern estuary – and showing the results to the public – may be coming to an end. Grid maps of the natural estuary pose a problem of giant proportions for Ecology's drive to implicate Capitol Lake – one so worrisome that they have once again "updated the model." This time around they have changed it so that it no longer shows any DO standards violations in the natural estuary at all (Figure 3-9). That would eliminate the complex method of calculating "violations" in the modern estuary – a good thing – but would also warp the model in ways that could make all of its predictions untrustworthy. That topic is analyzed in Chapter 10.

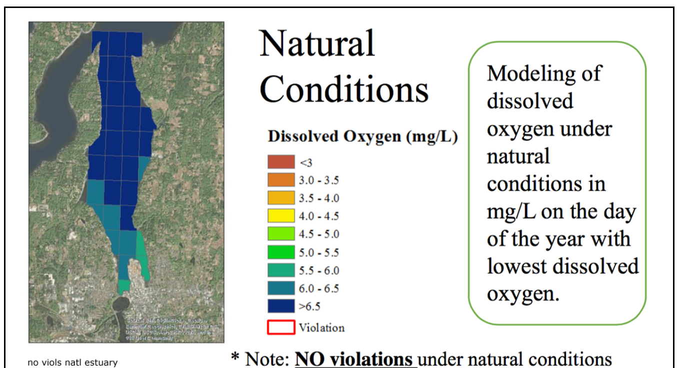

Figure 3-9. Slide from recent Dept. of Ecology presentation November 16, 2017 showing "natural" estuary with no DO violations anywhere, a result of a model change. Analyzed in Chapter 10, this Review. Source: Weiss, 2017.

[7] The model used here refers to all of Central and South Puget Sound, but is similar to the Budd Inlet model in its mode of calculation.