2015 Quality Control Report for Status and Trends Monitoring of Riparian and Aquatic Habitat in the Olympic Experimental State Forest

Prepared by Warren Devine and Teodora Minkova

January 13, 2016

Executive Summary

In 2012, Status and Trends Monitoring of Riparian and Aquatic Habitat on the Olympic Experimental State Forest (OESF) was implemented to document changes to riparian and in-stream conditions in basins managed by Washington State Department of Natural Resources (DNR) for timber, fish, and wildlife habitat and other ecosystem values. This report presents the results of a quality control (QC) analysis of 33 metrics that are monitored under the project. The objectives of this analysis were: (1) quantify the variability in the measurements of stream attributes within a field crew and between field crews, (2) quantify the between-year (annual) variability of monitoring metrics, and (3) provide recommendations for improvement of monitoring protocols, field training, temporal sampling design, and future status and trends analyses.

The metrics included in this analysis are all measured manually in the field and are therefore subject to various forms of measurement error that are not incurred by automated electronic sensors. The metrics assessed here quantify: stream morphology, channel coarse substrate, in-stream large wood, stream habitat unit classification, valley and channel classification, and riparian overstory vegetation. Field measurements were repeated during four visits to five sites to separately quantify the measurement error associated with the measurements themselves and that associated with interpretation of the protocols by different crews.

Overall, metrics of stream morphology, channel coarse substrate, valley and channel classification, and riparian overstory vegetation had acceptable levels of consistency, although we identify specific cases where clarifications in field protocols and modifications to field training could further improve measurement precision.

Metrics describing in-stream large wood had an undesirable level of measurement error, and this report recommends modifications to our field protocols and training to reduce the sources of error. We also found that our stream habitat unit classification metrics require improved field training to increase consistency among different crews.

For eight of our metrics, we were able to directly compare our measurement error with that reported for other regional status-and-trend stream habitat monitoring projects (Roper et al. 2010). Our levels of measurement error were similar to, or lower than, those of the other projects. Comparison of our QA/QC procedures for field protocols with those of the other monitoring projects led us to the conclusion that our QA/QC procedures are sufficiently rigorous given the project objectives, geographic scale, and budget.

Continuing to apply the QA/QC procedures and improving the field protocols as recommended in this document will be sufficient to achieve the desired data quality to characterize status and trends in aquatic and riparian habitat across the OESF. This will not only improve our confidence in the project findings but will allow sharing data with other projects for broader assessments at regional and national scale.

Table of Contents

This report describes the quality control (QC) process for field sampling and provides the results from the 2015 QC analysis for the ongoing project “Status and Trends Monitoring of Riparian and Aquatic Habitat in the Olympic Experimental State Forest”. The goal of the monitoring project is to evaluate changes in aquatic and riparian habitat over a broad landscape (≈270,000 ac or 110,000 ha) managed by Washington Department of Natural Resources with the dual objectives of revenue production and habitat conservation.

The focus of this report is QC of field protocols; thus it does *not* discuss:

1. Office data management quality assurance/quality control procedures, such as data verification and archiving. These are described in the monitoring protocols.
2. Statistical design of the overall monitoring project, which is described in the project’s study plan (Minkova et al. 2012) and establishment report (Minkova and Vorwerk 2014)

Need for QC of field protocols

The OESF riparian and aquatic habitat monitoring project tracks changes in environmental conditions resulting from ongoing management and natural disturbances (Minkova et al. 2012). The management footprint from timber harvest and roads in the OESF is (and is expected to continue to be) relatively small compared to previous decades and to industrially-managed lands. Therefore, we expect the resulting environmental changes in the monitored streams to be small. To detect these small changes, we need suitable sampling design, appropriate monitoring indicators and associated metrics, and sufficient sampling intensities. Equally important is our ability to identify the sources of variation and quantify them so that the trend analyses distinguish between changes resulting from environmental dynamics and those resulting from inconsistent methods and data collection.

In addition to contributing to future trend analyses, we will use the findings from this QC analysis to assess the performance of the habitat metrics identified in monitoring protocols and recommend changes to improve the field measurement procedures, to increase efficiency and lower cost of field work, and to improve in field crew training. Finally, by quantifying the variation in metrics between years and comparing to other sources of variation, we seek to inform the temporal study design for our project. For example, if a metric shows small year-to-year variation, sampling can occur less frequently than for metrics with greater variation over time.

QC Objectives

Several main reasons for measurement error has been identified and assessed in similar stratus and trends monitoring projects: operational definitions for the habitat attribute, field procedures, training and experience of the field crew, the intensity of measurements, and when and where the attribute is measured (Roper et al. 2010).

Our objectives are:

1. Quantify the amount of variability in the measurement of stream attributes used to monitor habitat conditions

We will report separately the variability within a field crew (“same-crew-and-year”- the same crew performing the same protocols twice in the same year) and between field crews (“between-crew”- different crews performing the same protocols in the same year). The first will inform us about the consistency (repeatability) of the field protocols, and the second will inform us about the inter-observer measurement error. Together, these will indicate what level of measurement error to expect for each measurement and whether modification of the field procedures and field training are needed.

2. Quantify the between-year (annual) variability of monitoring metrics

We will report differences between measurements collected by the same crew in 2014 and again in 2015. This will inform us about the magnitude of annual change to detect using selected metrics. Some of the channel geomorphology metrics do not typically change over such a short interval, which is why the monitoring protocols calls to re-measure them every five years. For those metrics, the QC analysis will test our assumption of no, or little, change after one year.

3. Provide recommendations for improvement of monitoring protocols (field procedures, work flow in the field, and operational definitions), field training, temporal sampling design, and future status and trends analyses.

Assessed field protocols and habitat metrics

Ten habitat monitoring protocols are currently implemented as part of Status and Trends Monitoring of Riparian and Aquatic Habitat on the OESF (Minkova and Foster in prep.).

Six of the ten protocols are evaluated in this QC analysis. Five of these protocols are stream survey protocols: stream morphology, coarse substrate, in-stream large wood, habitat units, and valley and channel classification. The sixth protocol is riparian vegetation.

Monitoring protocols not included in this QC analysis are: stream shade, stream temperature, stream discharge, and riparian microclimate. Stream shade is seasonally sensitive and could not be assessed in 2015 because some of our QC field visits occurred after leaves had begun to fall. Stream shade QC sampling will be conducted during summer 2016. Most of the data for the stream temperature, stream discharge, and riparian microclimate protocols are recorded by automated sensors and are not included in this QC analysis. Separate QC processes have been developed and implemented for each of them. They are described in the monitoring protocols and in the 2014 progress report (Minkova and Devine 2015).

All measurements for this QC analysis are taken using the definitions and following the field procedures described in the stream survey field protocols and the riparian vegetation protocol monitoring protocols (Minkova and Foster in prep.). A list of all the metrics included in this QC analysis appears in Table 1.

Table 1. List of metrics included in the QC analysis.

Metric	**Explanation**
Stream Morphology	
Channel gradient	Percent slope of the water surface, calculated using total cumulative difference in water surface elevations (cross-sections A through F) and total reach length.
Bankfull width	Average of bankfull width measured at each of the six cross sections.
Bankfull depth	Mean bankfull depth, averaged across all six cross sections. At each cross section, mean bankfull depth is calculated from the bankfull depths at the 11 stations between the bankfull stations. At each station, depth is calculated as the mean of the two bankfull elevations minus the elevation of the streambed at that station.
Floodplain width	Mean floodplain width, measured at three cross sections per sample reach.
Bankfull width: depth ratio	Ratio calculated for each cross-section and then averaged by sample reach.
Bankfull thalweg depth	An average of the maximum bankfull depth at each of the six cross sections. Maximum bankfull depth at each cross section is calculated as the mean of the two bankfull elevations minus the elevation of the streambed at the thalweg.
Erosion	Percentage erosion is calculated as the sum of streambank distance that is actively eroding along both sides of the sample reach, divided by twice the sample reach length, then multiplied by 100. It should be noted that the sample reach length is measured along the thalweg and thus will likely differ from the length measured along each bank.

Metric	**Explanation**
Channel Coarse Substrate	
D_{50} (median particle size)	Particle size was recorded categorically (with the exception of some larger particles that were subsequently categorized based on measured diameter), so D_{50} is derived by ordering the categories, calculating the cumulative percentage of samples in each category, and then identifying the category containing the median value. Size categories followed those of the gravelometer, with "<2 mm" defining the smallest size class (including sand, silt, and clay). All particles between 250 and 3999 mm were put into the "boulder" class (particles in this size range were sometimes measured or sometimes simply recorded as "boulder" in the field). Organics and bedrock were not included in the D_{50} calculation.
Substrate percentage by size class	Percent of substrate in each particle size class (Kaufmann et al. 1999), as listed in the protocol: Bedrock Boulders (>250 to 3999 mm) Cobbles (>64 to 250 mm) Gravel (coarse) (>16 to 64 mm) Gravel (fine) (>2 to 16 mm) Sand and finer (2 mm and smaller, including sand, silt, clay, soil) Organic (all types of organic materials)
Percent fines	Percentage of all particle size samples in the sand and finer class (2 mm and smaller, including sand, silt, clay, and soil)
Embeddedness	Mean embeddedness of particles within the size classes for which embeddedness was assessed (boulder, cobble, and gravel (coarse)). Note that embeddedness was not assessed for all of the particles in the gravel (coarse) class.
In-Stream Large Wood	
Total pieces/100 m (excluding pieces in jams)	Count of total pieces (logs and rootwads) in sample reach, adjusted to 100-m length. Pieces in jams are not included.
Total pieces/100 m (including pieces in jams)	Count of total pieces (logs and rootwads) in sample reach, adjusted to 100-m length. Pieces in jams *are* included. This metric was added because surveys sometimes disagreed on whether to count a jam. Such disagreements could significantly affect piece counts.
Percentage of pieces per species class	Percentage of pieces in each species class (conifer, deciduous, or unknown).
Percentage of pieces per decay class	Percentage of pieces in each decay class (1-5).
Piece mean diameter	Mean diameter of all pieces in the sample reach.
Total length of all pieces, **by zone**, per 100 m	Sum of piece lengths for all pieces in each zone, adjusted to a 100-m sample reach length. This metric was chosen instead of mean piece length because it accounts for missed pieces.
Pool-forming pieces per 100 m	Average number of pool-forming pieces per sample reach, adjusted to 100-m sample reach length.
Sediment-storing pieces per 100 m	Average number of sediment-storing pieces per sample reach, adjusted to 100-m sample reach length.
Piece orientation	Average percentage of pieces in each orientation class (A, B, C, D, vertical).

Metric	Explanation
Piece stability	Average percentage of pieces in each stability class (pinned, buried, rooted, unstable or stable).
Number of jams	Number of jams per sample reach.
Number of pieces in jams	Total number of pieces tallied in all jams in each sample reach.
Habitat Units	
Units/100 m	Total number of habitat units per sample reach, adjusted to a 100-m length.
Pools/100 m	Total number of pools (scour pool, dammed pool, backwater pool) per sample reach, adjusted to a 100-m length.
Number of habitat units/100 m, by type	For each habitat unit type, the average number of units per sample reach calculated, adjusted to a 100-m length.
Percentage of sample reach allocated to each habitat unit type	Average percentage of sample reach allocated to each habitat unit type.
Residual pool depth	For each pool, residual pool depth is calculated as the maximum depth minus the tail-crest depth. Residual pool depth is then averaged across all pools.
Valley and Channel Classification	
Valley segment and channel type	Valley segment classification and channel type classification are compared for each sample reach.
*Riparian Vegetation (Overstory)**	
No. trees per basin (2 sample plots)	Total number of trees counted in the two sampling plots in each basin.
Trees/ha by plot	Number of trees per hectare for each plot in each sample reach.
Basal area/hectare by plot	Basal area per hectare on each plot in each sample reach.
Percentage of trees by species per basin	Percentage of trees by species per basin.

*QC analyses were not conducted on the understory and canopy closure measurements in riparian vegetation protocol. The dimensions of the sample plots were not confirmed, but the consistency of plot position per the protocol was noted.

Monitoring sites used in the QC analysis

The status and trends monitoring project monitors 50 type-3 basins (drainages of the smallest fish-bearing streams) in the OESF and 4 reference basins of comparable size in the Olympic National Park. The OESF sample will be analyzed for trends and management effects. The sites in the park will be used mainly to give a picture of how the aquatic and riparian conditions change by natural disturbances only. Field sampling takes place in a permanently marked stream reach at the outlet of each basin (the most downstream segment of the Type-3 stream). These sample reaches have length of 20 bankfull widths or at least 100 m.

By the end of the 2015 field season, the sample reaches of all 50 monitored OESF basins were sampled following the stream survey protocol. Five of the 50 OESF basins (10%) were selected for QC. This amount of QC effort is comparable to the U.S.D.A. Forest Service's Aquatic and Riparian Effectiveness Monitoring Plan (AREMP) QC protocol where they re-sample 10% of the basins for inter-observer variability (Lanigan 2014).

The riparian vegetation monitoring protocol was implemented in 40 of the 50 OESF monitored basins in 2014 and 2015. Four of those 40 basins (a 10% sample) were selected for QC analysis.

Selection of the QC sites

The following criteria were used in order to select five basins for QC of the stream survey protocols:

- Basins initially sampled in 2014 by the main field crew – to quantify between-year variability (2014 – 2015)
- Basins accessible and with enough surface flow for sampling in 2015 - to quantify between-year variability, to conduct a second 2015 visit by the same crew to quantify the same-crew-and-year variability, and to conduct a third 2015 visit by a different crew to quantify the between-crew variability
- Basins sampled at approximately the same time in 2014 and 2015 – to minimize the effect of seasonal variation, which could potentially impact these survey protocols through seasonal differences in stream flow and stream-bank vegetation.
- Basins without major management disturbances (such as timber harvests and road management near the sample reach) and without large natural disturbances (such as windthrow and landslides) in 2014 and 2015 – to minimize the influence of site-specific variability.

The selection criteria for the four riparian monitoring sites were different since the initial sampling was done by various crews over the two years and they were not available for repeat visit. Thus, it was not possible to assess the same-crew-and-year and between-year variability for the riparian vegetation overstory. Only the between-crew variability was assessed. All 4 basins had to be initially sampled by a single crew earlier in 2015, to be accessible for sampling in 2015 and to no be subject to major management or natural disturbances.

Five basins met the selection criteria QC stream survey sampling: 158, 488, 718, 724, and 763. The four basins sampled for riparian vegetation QC were 158, 584, 718, and 763 (Figure 1).

Field crews

The project main field crew for stream survey protocols consisted of Ellis Cropper (ESC) and Mitchell Vorwerk (MV). These two crew members performed most of the field work in 2014 and 2015 and have been trained at the beginning of each field season. They revisited the five selected basins that needed re-measurements to assess the same-crew-and-year variability and the between-year variability.

A different crew (Alex Foster and Teodora Minkova), who are the principle researchers on the project and authors of most monitoring protocols, visited the same five basins in 2015 to collect data to assess between-crew variability.

A DNR crew consisting of Richard Bigley (the researcher overseeing the riparian vegetation monitoring and author of the riparian vegetation monitoring protocols) and Warren Devine (OESF data manager) re-measured the paired riparian vegetation overstory plots at four basins.

Sampling period

All QC visits were conducted between August 18 and September 24, 2015. The period between repeated visits of the same crew was minimum one week. The crews did not work in the same basin at the same time.

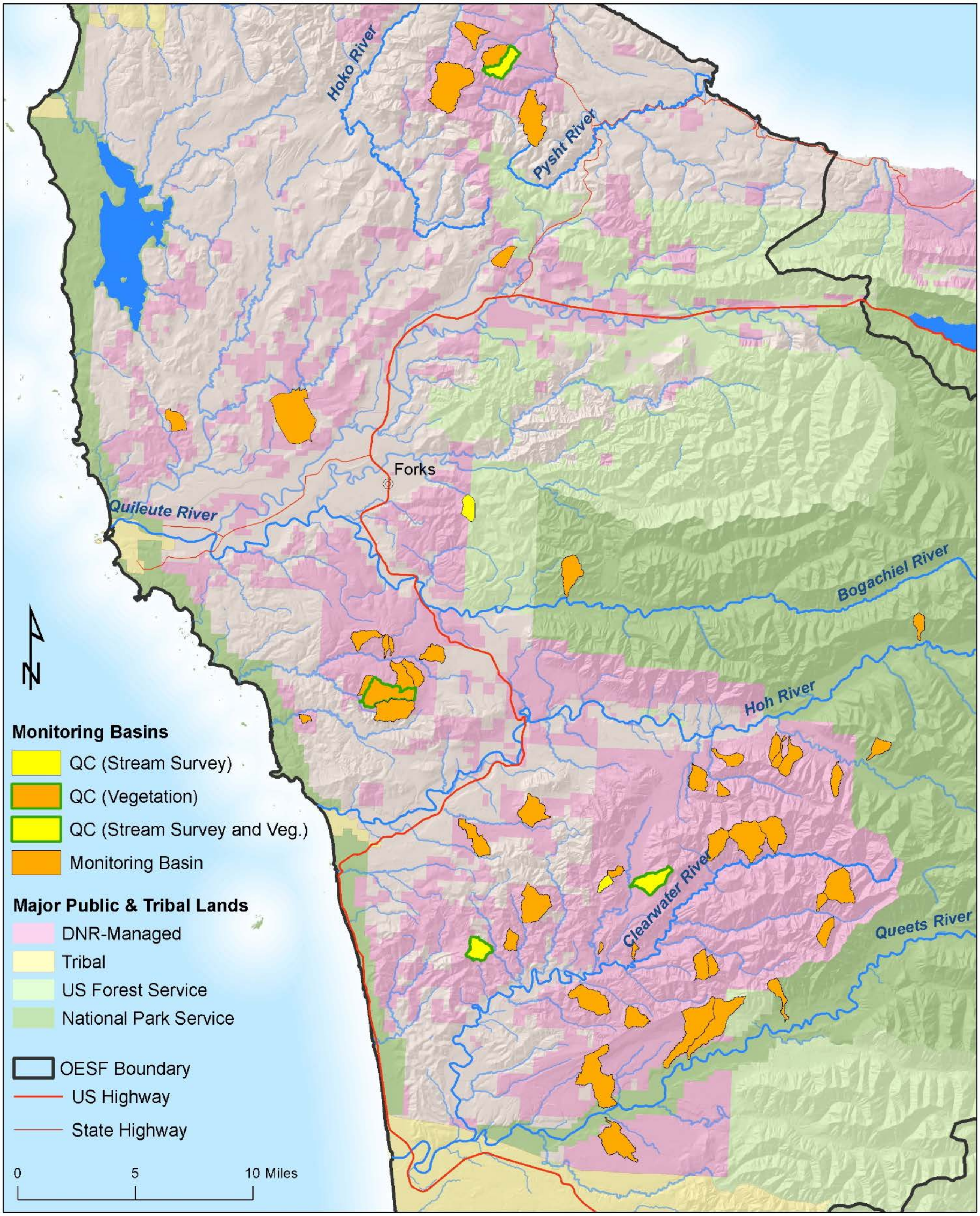

Figure 1. Monitored basins in the OESF and Olympic National Park. Stream Survey QC basins are shown in yellow and Overstory Vegetation QC basins are outlined in green.

QC analyses

General Approach

The first objective of this QC analysis was to assess between-crew variability and same-crew-and-year variability. This was achieved using the following comparisons:

1.a. Between-crew variability: Data collected by the MV, ESC crew was compared with that of the TM, AF crew collected in the same year.

1.b. Same-crew-and-year variability: The MV, ESC crew visited the same five sample reaches twice in 2015 and data collected on these two visits were compared.

The second objective of this QC analysis was to assess between-year variability, in this case variability associated with the passage of one year (2014 to 2015). This was assessed by:

2. The MV, ESC crew visited the same five sample reaches in 2014 and 2015; the measurements were compared to assess temporal variability. It should be noted that this comparison includes both temporal variability *and* intra-observer variability. However, by considering the magnitude of the 2014 vs. 2015 comparison in light of the magnitude of the intra-observer comparison, we can draw general conclusions about the relative amount of temporal variability.

These three sources of variability were assessed using three different datasets and models (described below). Owing to the fact that they were separate datasets and models, we cannot directly compare the magnitude of variability among the between-crew, same-crew-and-year, and between-year factors.

The only way to include all three comparisons in the same model and dataset would be to use a full factorial sampling design for QC: sampling in two years, with both crews re-visiting all five selected basins twice in each year. In such a design, year, crew, and re-visit would all be factors with two levels each. That type of analysis could be conducted in the future, but cannot be used at present given the dataset that we have in this first year of QC.

Analysis

We used mixed-effects analysis of variance (ANOVA) to test the significance of our three comparisons and to partition the variability in the data (variance decomposition). In this model, site was a random (block) effect and the comparison factor (visit, crew, or year) was a fixed effect (Table 2). Thus, the models test the magnitude of each fixed-effect variability source (between-crew, same-crew and-year, and between-year) relative to unexplained variability (residual error), while accounting for the natural differences among sites. To help visualize the sources of variance, we use Eta^2 to calculate the proportion of the sum of squares associated with each source (Brown 2007).

Table 2. Analysis of variance (ANOVA) models used to analyze each QC metric and to calculate the proportion of variance associated with each source (Eta2).

Source	df	Sum of Squares	Mean Square	F value	Eta^2
Same-crew-and-year variability model					
Site	4	SS_b	$SS_b/4$		SS_b/SS_t
Visit[a]	1	SS_c	$SS_c/1$	MS_c/MS_e	SS_c/SS_t
Residual Error	4	SS_e	$SS_e/4$		SS_e/SS_t
Total	9	SS_t			
Between-crew variability model					
Site	4	SS_b	$SS_b/4$		SS_b/SS_t
Crew[b]	1	SS_c	$SS_c/1$	MS_c/MS_e	SS_c/SS_t
Residual Error	4	SS_e	$SS_e/4$		SS_e/SS_t
Total	9	SS_t			
Between-year variability model					
Site	4	SS_b	$SS_b/4$		SS_b/SS_t
Year[c]	1	SS_c	$SS_c/1$	MS_c/MS_e	SS_c/SS_t
Residual Error	4	SS_e	$SS_e/4$		SS_e/SS_t
Total	9	SS_t			

[a] Compares MV, ESC 2015 visit 1 with MV, ESC 2015 visit 2.

[b] Compares MV, ESC 2015 visit 1 with TM, AF 2015 visit.

[c] Compares MV, ESC 2014 visit with MV, ESC 2015 visit 1.

The majority of the metrics analyzed in this report produce a single numerical value (e.g., sample reach gradient or percent bank erosion). For these metrics, we use an analytical approach that includes the following:

1. Variance decomposition, presented in a horizontal bar graph, and calculated using Eta^2 as shown in the three ANOVA models in Table 1.
2. Tests of significance (F-tests) for the Visit, Crew, and Year factors, using the three ANOVA models in Table 1. For the sake of brevity, we do not include the ANOVA tables in our results. Instead, we simply state when any of the three factors (visit, crew, or year) are statistically significant (defined here as $P < 0.05$).
3. Evaluation of the difference between measurements. In the comparisons, neither measurement is considered more accurate than the other. We are interested in how close the two measurements are to one another, regardless of which direction each measurement deviates. Therefore, the differences are expressed as absolute values. The absolute value of the difference is also expressed as a relative magnitude (%), calculated as:

   (measurement A – measurement B) / (measurement A + measurement B / 2)

   An exception to this calculation method was made for the riparian overstory analysis. The overstory measurements were made by crews with contrasting levels of experience (novice summer crews vs. the experienced DNR QC crew). Thus, we regard the DNR crew's data as "truth" and the differences in measurements were attributed to error by the summer crew. We report the error as positive if the summer crew reported a greater value and negative if the summer crew reported a smaller value, relative to the DNR QC crew. The percentage error is calculated as:

   (Summer crew measurement – DNR crew measurement) / (DNR crew measurement)

Some metrics do not produce a single numerical value: some are categorical (e.g., valley segment type) and some are based on a percentage distribution of samples among classes (e.g., percentage of sample reach in each habitat unit). For these variables, the data are presented in simple bar graphs or in tables.

In the discussion section at the end of this report, we present a table of signal-to-noise ratios (S:N) for all of the metrics that were analyzed using continuous variables (i.e., not metrics that were analyzed categorically such as decay class of LWD or substrate particle size distribution). We calculated S:N following the procedure from Roper et al. (2010): "a random-effects ANOVA model was used to decompose the total variance into that associated with differences among streams versus variation in crew observations at a stream (all error not due to the main effect of stream site is treated as observer variability…)". Thus, we used the sum of squares from the ANOVA and calculated S:N as:

$$(SS_b) / (SS_c + SS_e)$$

Where SS_b is the sum of squares associated with site, SS_c is the sum of squares associated with the comparison effect, and SS_e is the residual error sum of squares. S:N was calculated for two comparisons: the between-crew comparison and the same-crew-and-year comparison.

Interpreting the Results

Interpretation of our variance distribution horizontal bar graphs is somewhat different from the interpretation of those published for other projects (e.g., Larsen et al. 2004). In ours, each set of three bar graphs represents the *same* dependent variable, and the different graphs are associated with different comparisons (i.e., visit, crew, or year).

For metrics analyzed using variance decomposition, we present the differences between measured values in each comparison using two vertical bar graphs. The bars represent the means of the absolute value of the differences between metrics (n=5, because there are 5 sample reaches). The left bar graph (green) shows the difference in measurement units. The right graph (yellow) shows the relative magnitude of the difference (%). The black dots indicate the minimum and maximum absolute differences among the five sample reaches.

In interpreting each set of results graphs (the variance distribution graph and the two vertical bar graphs), the variance distribution graph shows how much of the overall variance is explained by site, how much is explained by the comparison in question, and how much is associated with other unidentified sources (i.e., residual error). The ratio of the variance explained by the comparison to the amount of residual error is used to calculate the statistical significance of the comparison (Table 2).

The magnitude of the differences shown in the green and yellow bar graphs is not necessarily correlated with the size of the orange bar. An orange bar is relatively large where there is a difference between the two levels of the comparison factor *and* that difference is consistent. The orange bar may be very small if there is little difference between the two crews' measurements *or* if the differences between the two crews' measurements is inconsistent and averages to zero or near zero.

Examples of three potential outcomes of the QC analysis, using the between-crew comparison

1. *The difference is not statistically significant, and the values represented by the green and yellow bars are large.* In this case, the two crews got much different values for their measurements in all five sample reaches, but there was no consistent pattern to the differences (i.e., no detectable bias). In summary, the crews could not measure the metric consistently, and it was probably not a matter of protocol interpretation because the measurement error was not consistent. The protocol itself should be re-assessed for definitions of the habitat attributes, repeatability of the field procedures etc.
2. *The difference is not statistically significant, and the values represented by the green and yellow bars are small.* The small green and yellow bars indicate that the measurements of the two crews were consistent. This suggests that the protocol, and the execution of the protocol in the field, were both successful.
3. *The difference is statistically significant, regardless of the magnitude of the values represented by the green and yellow bars*. In this case, one crew got a measurement that was consistently greater than that of the other crew across all five sample reaches. This could result from protocol misinterpretation or from an instrument calibration error. Further evaluation is needed to detect the source of the bias.

The two S:N statistics summarized in the discussion section have different interpretations. The same-crew-and-year S:N statistic is an indication of how precise a metric is, under ideal conditions: the same trained and experienced field crew applying their interpretation of the same protocol to the same stream on two different visits in the same year. The between-crew S:N statistic indicates the precision of a metric under more typical circumstances: multiple trained crews interpreting and applying the same protocol. Roper et al. (2010) suggests guidelines for interpreting S:N for stream channel metrics: "…we characterize the likelihood of detecting environmental heterogeneity [i.e., detecting real differences] as high when S:N ratio is greater than 6.5, moderate when S:N ratio is between 2.5 and 6.5, and low when S:N ratio is less than 2.5."

Results

Field Effort

The QC protocols were completed in the same amount of time as the original sampling: approximately 2 days per basin including travel time. Thus, the QC protocols were completed in 20 additional field days for the MV, ESC crew and 10 additional field days for the TM, AF crew to conduct stream survey protocols in 5 sample reaches. It took 4 field days (one day per basin including travel time) for the RB, WD crew to resample riparian vegetation overstory.

Results by Metric

QC results for each of the 33 aquatic and riparian habitat metrics are reported below.

CHANNEL GRADIENT

- The proportion of variance associated with the three comparison effects was negligible.
- There were no significant differences in any of the three comparisons.
- Overall, gradient measurements were very consistent within each of the three comparisons, with differences averaging less than 0.4 percent slope.
- The magnitude of the between-crews variance is primarily due to a single, relatively large, calculation error made in 2015 on one of the paper field forms.
- Much of the between-years variance is associated with a single measurement, apparently erroneous, made in 2014. It is not clear what the cause of the error was.

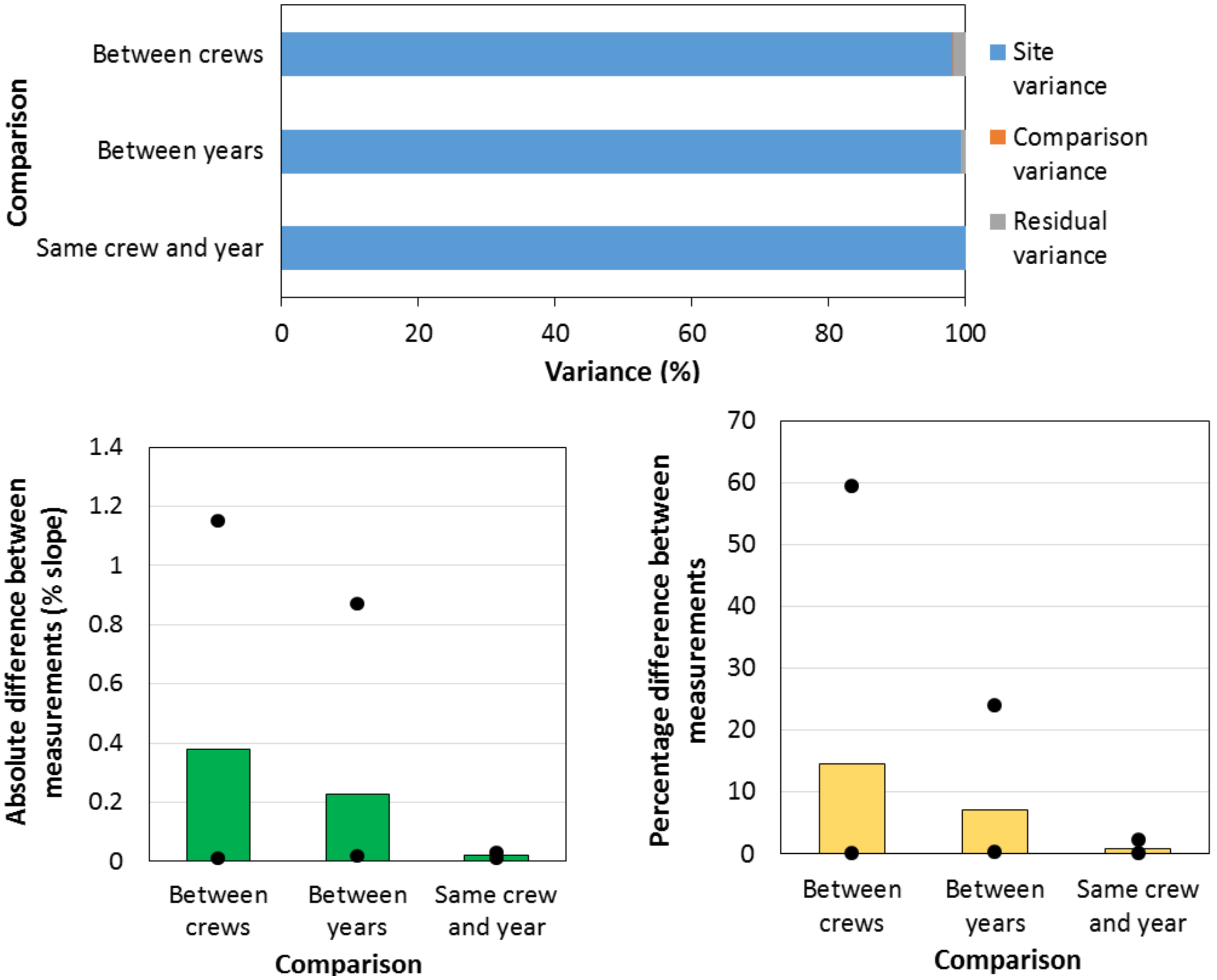

Figure 2. Variance distribution (top), measurement differences (lower left), and percentage measurement differences (lower right) for three comparisons of channel gradient measurements made as part of a QC assessment. Green and yellow bars represent means (n=5); black dots represent maximum and minimum values.

Conclusion: The field protocol and training are adequate, given small difference of <10 % (the two data points above that value were due to calculation error and to a single erroneous measurement or data entry mistake). All three differences were less than 1% slope, which is the resolution of a clinometer. This supports the researchers' decision to use auto level for more precise measurement of gradient.

Recommendation: Elevation differences should be calculated automatically in the office, not manually in the field, to reduce the probability of calculation error.

BANKFULL WIDTH

- There was no significant measurement difference associated with crew, year, or visit within the same crew and year.
- Overall, bankfull width measurements were relatively consistent, with measurement differences averaging less than 10% for all three comparisons.
- After averaging the six cross-sections by sample reach, there were no obvious measurement outliers. The greatest difference among measurements at an individual cross-section was 2.92 m (4.0 vs. 6.92 m), which occurred between different crews in the same year.

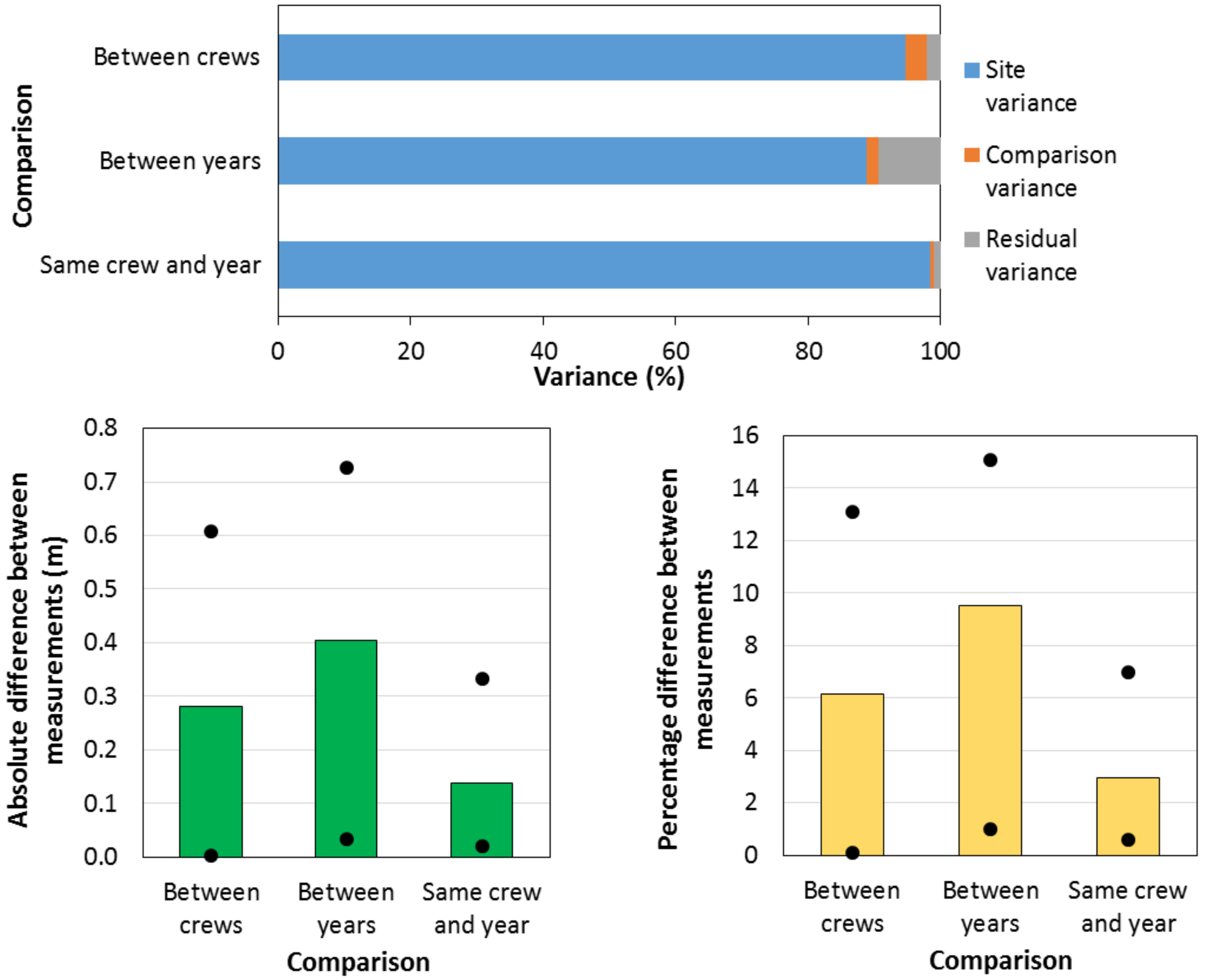

Figure 3. Variance distribution (top), measurement differences (lower left), and percentage measurement differences (lower right) for three comparisons of bankfull width measurements made as part of a QC assessment. Green and yellow bars represent means (n=5); black dots represent maximum and minimum values.

Conclusion: The field protocol and training are adequate: all three difference are <10 %. The consistency between and within crews is higher than expected given the complex stream banks in the sample reaches and the multiple qualitative indicators of bankfull stage described in the monitoring protocol.

Recommendation: For each new crew, conduct comprehensive training focusing on consistent identification of bankfull stage indicators. If the field crew is the same, continue to conduct annual calibration field training. Specify a method of lining up the top and bottom tapes at a cross-section .

BANKFULL DEPTH

- Measurement differences between crews averaged less than 3 cm; there was no significant difference associated with crew.
- Measurement differences between years 2014 and 2015 averaged 4.3 cm (23%), a significant difference.
- Differences between measurements by the same crew in the same year averaged less than 2 cm, not a significant difference.

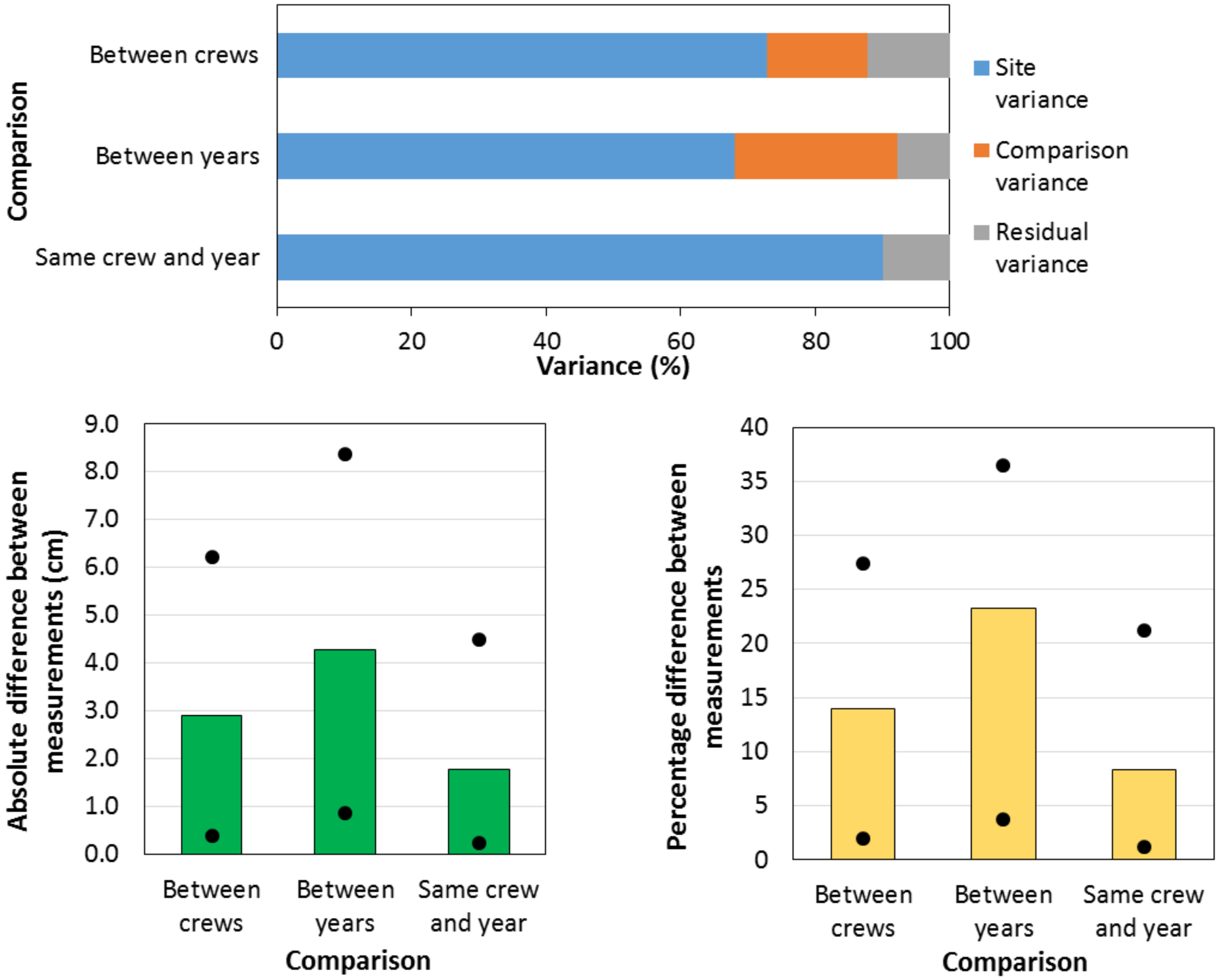

Figure 4. Variance distribution (top), measurement differences (lower left), and percentage measurement differences (lower right) for three comparisons of bankfull depth measurements made as part of a QC assessment. Green and yellow bars represent means (n=5); black dots represent maximum and minimum values.

Conclusion: Given the dynamic nature of the small streams sampled in the OESF, a difference of 4 cm between years is not unexpected. For example, a winter high flow could erode the bank and make it difficult to choose the same bankfull stage next year. A potential explanation for the 14% mean differences between crews is that relatively small differences in the identified bankfull stage can introduce relatively large differences in bankfull depth measurements.

Recommendation: Accept a relatively large margin of error for this metric in future trend analyses. Owing to the difficulty of consistently identifying bankfull stage, we should have intensive pre-season field calibration for crew members, especially new crew members, in multiple basins. This is important because bankfull indicators can be very different between basins (different flow regimes, bank substrates, etc.).

BANKFULL WIDTH: DEPTH RATIO

- Bankfull width-to-depth ratio was calculated for each cross-section and then averaged by sample reach. The mean ratio across all sample reaches and surveys was 24.1: 1.
- There were no significant differences among the three comparisons.

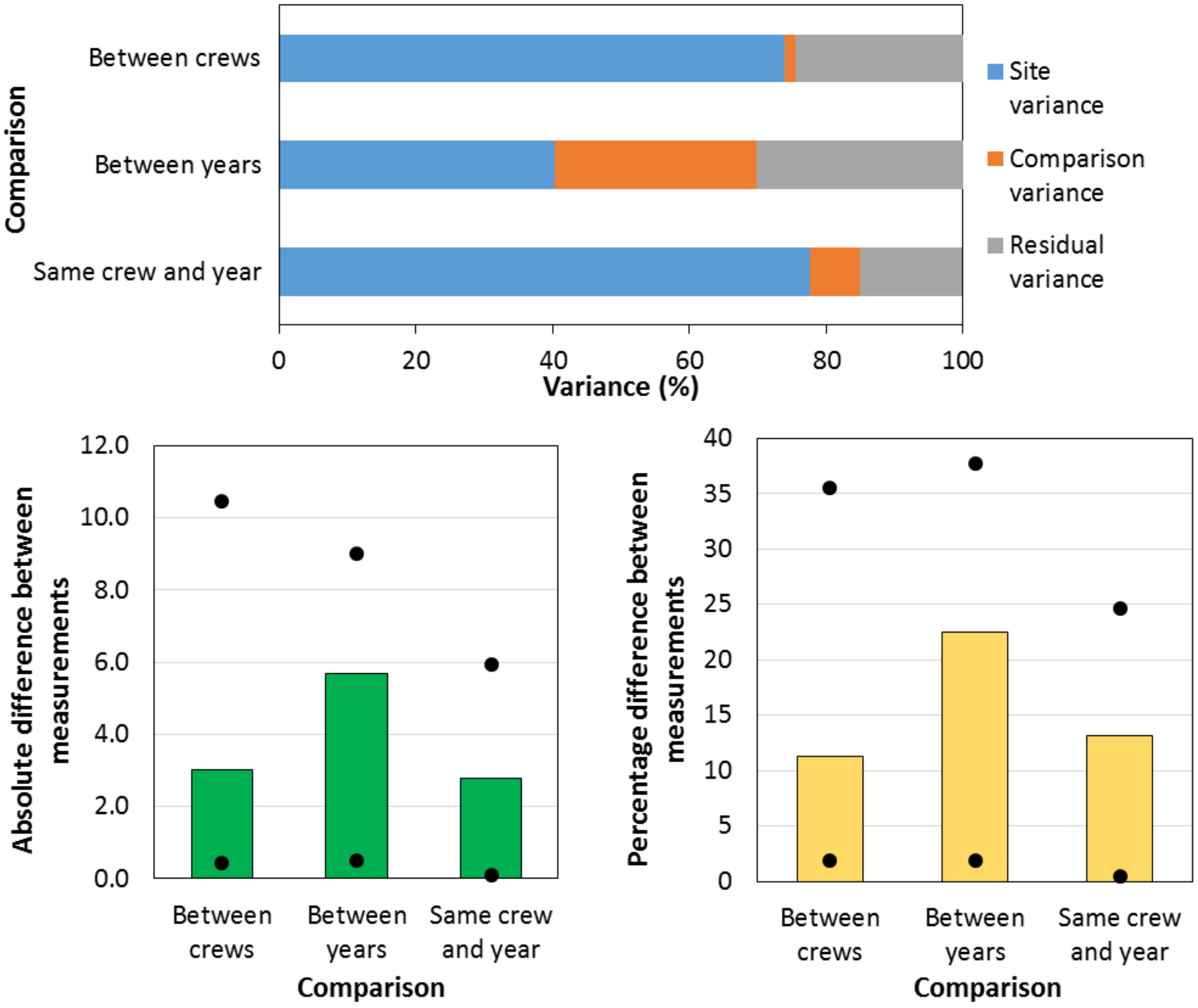

Figure 5. Variance distribution (top), measurement differences (lower left), and percentage measurement differences (lower right) for three comparisons of bankfull width: depth ratio measured as part of a QC assessment. Green and yellow bars represent means (n=5); black dots represent maximum and minimum values.

Conclusion: The observed differences are result of the interaction of bankfull width and depth measurements discussed above.

Recommendation: Follow the recommendations for bankfull width and depth measurements.

BANKFULL THALWEG DEPTH

- Differences between crews averaged 4.2 cm (11%), though this difference was not significant.
- Bankfull thalweg depth averaged 5.4 cm (15%) greater in 2015 than in 2014, a significant difference. The greatest between-year difference at an individual cross-section was 52.25 cm (sample reach 763, cross section C). All three 2015 measurements at that cross section were much greater than the 2014 measurement. There does not appear to be any error in the measurements.
- Differences between measurements by the same crew in the same year averaged 2.5 cm (not significant).

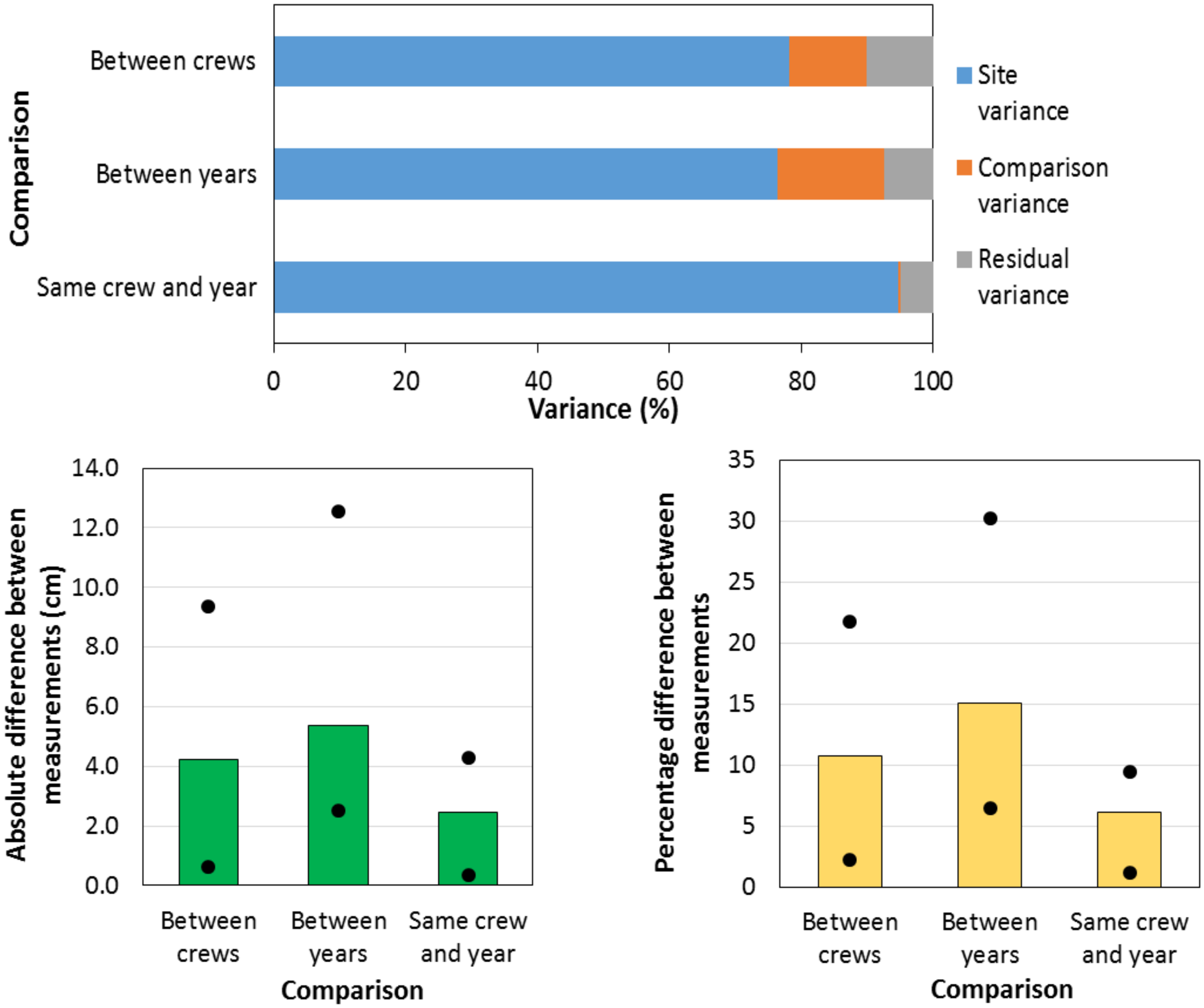

Figure 6. Variance distribution (top), measurement differences (lower left), and percentage measurement differences (lower right) for three comparisons of bankfull thalweg depth measurements made as part of a QC assessment. Green and yellow bars represent means (n=5); black dots represent maximum and minimum values.

Conclusion: Given the dynamic nature of the small streams in the OESF, a difference in thalweg depth of 5 cm between years (based on only 6 measurements per reach) is not unexpected. As with bankfull depth (above), the identification of bankfull stage markers may contribute to measurement differences.

Recommendation: If the field crew changes, conduct comprehensive field training on this protocol. If the field crew is the same, continue the current calibration field training. Focus training on identifying bankfull indicators in different type streams.

FLOODPLAIN WIDTH

- Floodplain width measurements differed significantly between crews; with the MV, ESC crew averaging 1.5 m greater than the TM, AF crew.
- Floodplain width measurements differed by an average of 1.4 m between years, but differences were not consistently in the same direction and were non-significant.
- Measurements within the same crew and year differed by an average of only 0.7 m (not significant).

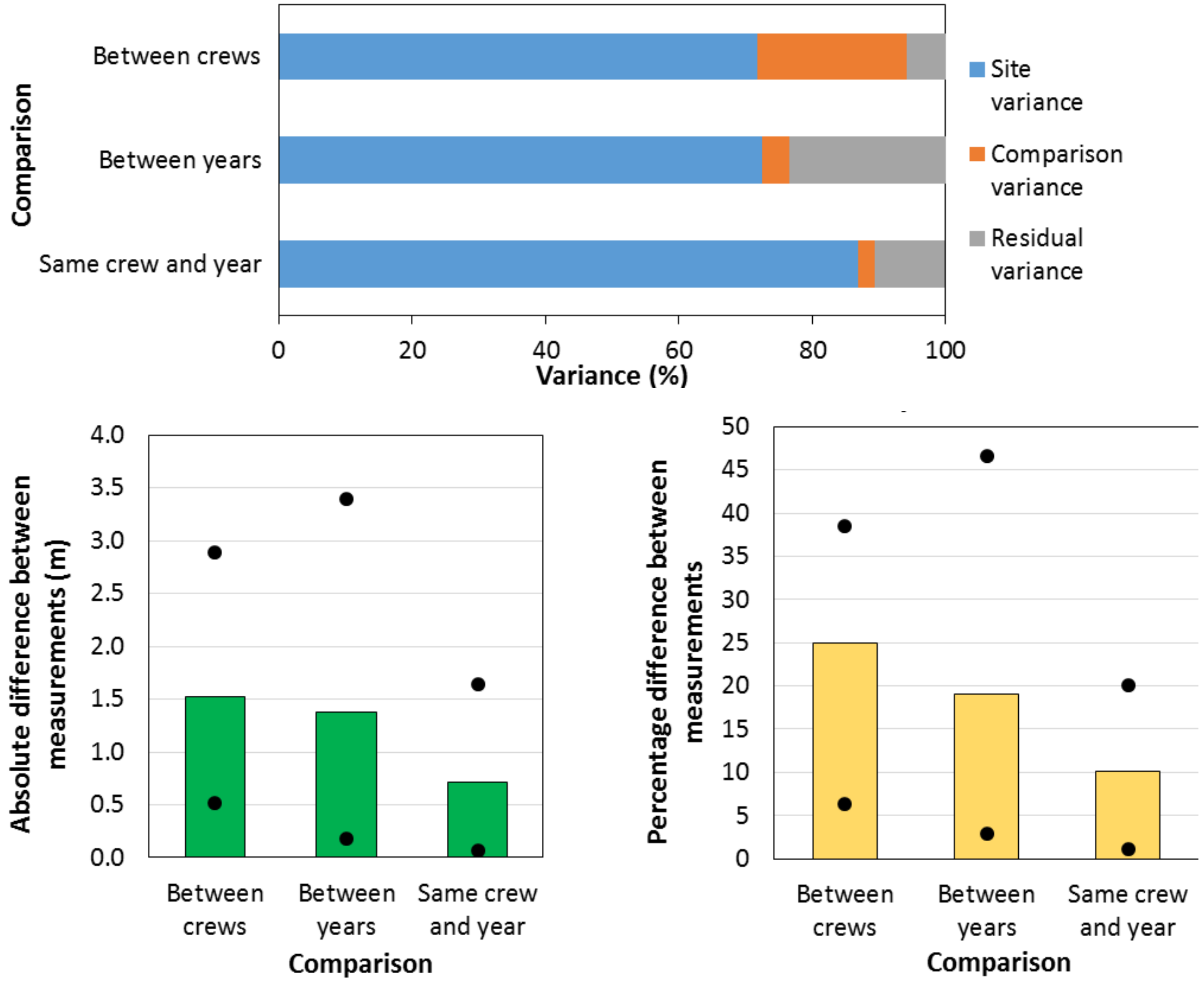

Figure 7. Variance distribution (top), measurement differences (lower left), and percentage measurement differences (lower right) for three comparisons of bankfull thalweg depth measurements made as part of a QC assessment. Green and yellow bars represent means (n=5); black dots represent maximum and minimum values.

Conclusion: Further investigation showed that the two crews used slightly different field procedures (TM, AF crew used only the doubled thalweg depth, while the ESC, MV crew primarily focused on topographic breaks on both banks as indicators of the floodplain. Given that there was no major flood event between 2014 and 2015, the difference between years is unexpectedly high. The overall higher variability of this metric may be due to the small number of measurements (3) taken per sample reach.

Recommendation: Improve the definition of floodplain and the field procedure in the monitoring protocol. If the field crew changes, conduct comprehensive field training on this protocol. If the field crew is the same, continue the current calibration field training.

EROSION

- There was a significant difference in erosion measurements between crews. The ESC, MV crew measured total streambank erosion, on average, 11 percentage points greater than the TM, AF crew (30% vs. 19% actively eroding bank).
- Differences in measured erosion between years and within the same crew and year were 7 percentage points or less (non-significant).

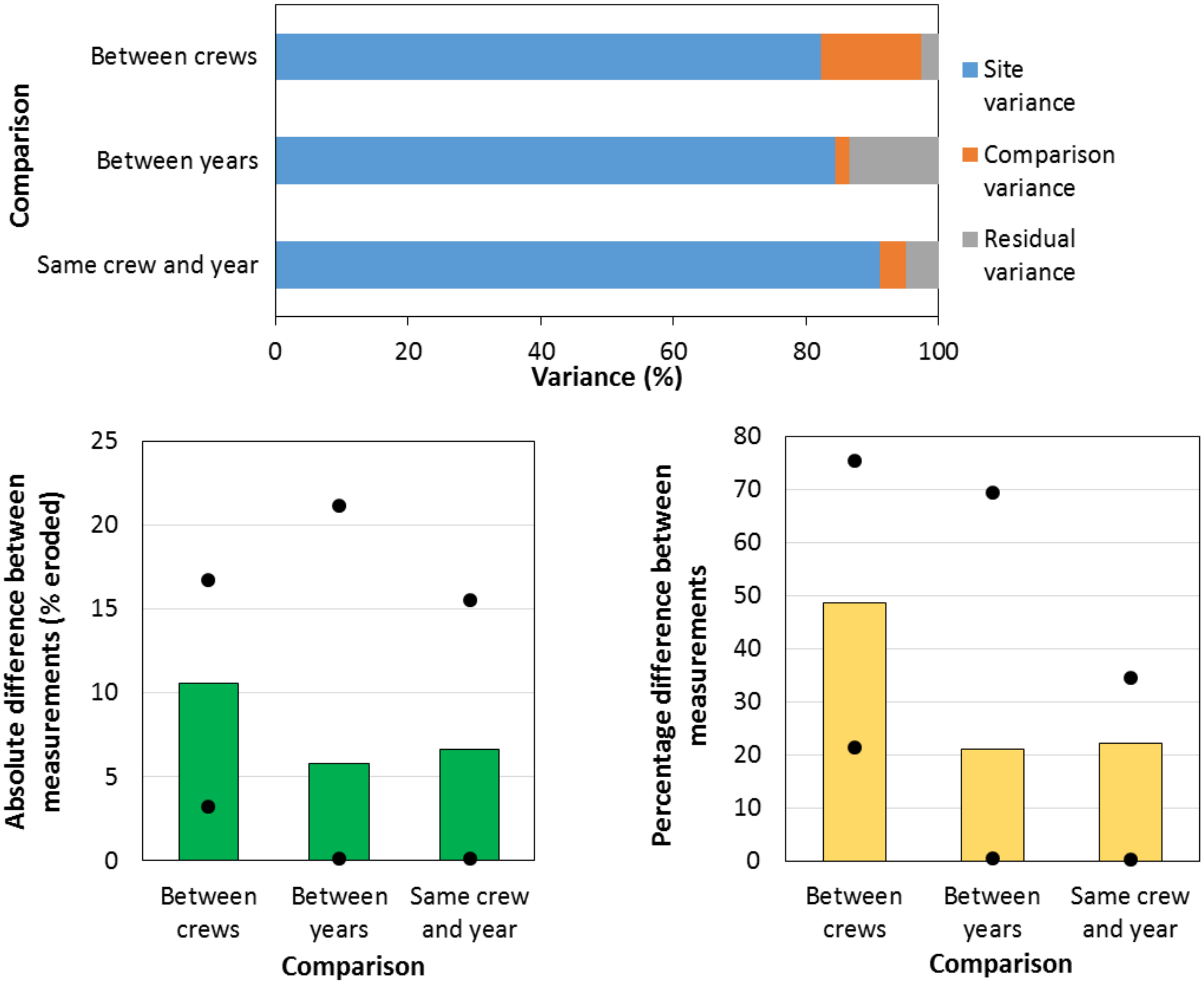

Figure 8. Variance distribution (top), measurement differences (lower left), and percentage measurement differences (lower right) for three comparisons of erosion measurements made as part of a QC assessment. Green and yellow bars represent means (n=5); black dots represent maximum and minimum values.

Conclusion: This metric is inherently variable as evident from the same-crew-and-year variability. However, this doesn't entirely explain the large (49%) between-crew variability. Further investigations showed the field crews were not clear on the criteria what qualifies as *active* erosion patch. In addition, one of the crews was estimating the length or was measuring the length less precisely than the other.

Recommendation: Accept a large margin of error for this metric in future trend analyses. Improve the description of qualifying active erosion patch in the monitoring protocol. Enhanced field training may additionally improve the precision. Consider other field protocols, although the reports from different protocols report mixed success and attribute this to the subjective nature of most protocols (Archer et al. 2004)..

COARSE CHANNEL SUBSTRATE: D_{50}

- There was no significant difference in D_{50} for any of the three comparisons.
- Between crews, the greatest discrepancy in D_{50} occurred in sample reach 158 (180 vs. 64 mm)
- Between years, the greatest discrepancy in D_{50} was in sample reaches 724 and 763 (both 32 vs. 45 mm)
- Within year and crew, the greatest discrepancy in D_{50} was in sample reach 724 (32 vs. 45 mm)

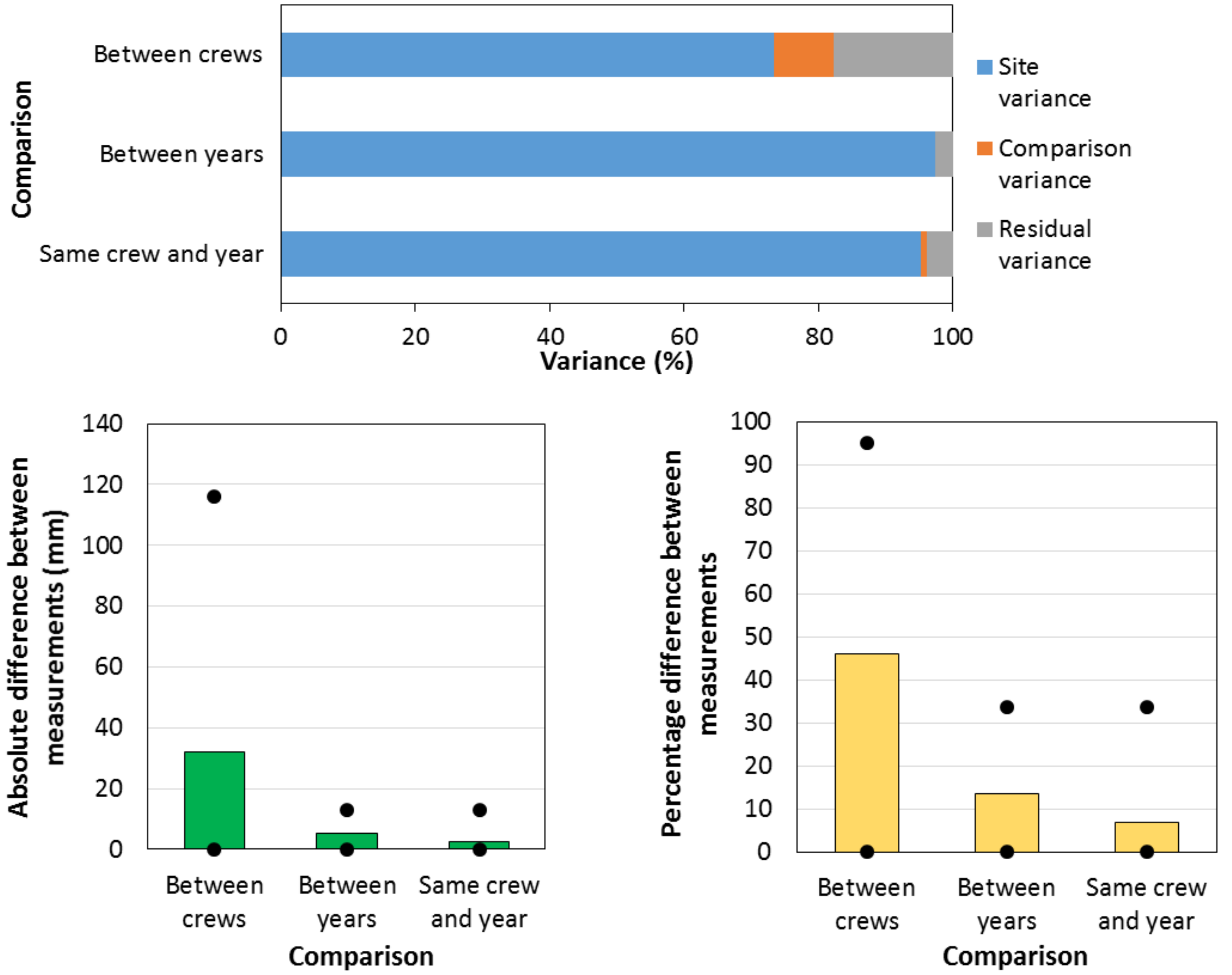

Figure 9. Variance distribution (top), measurement differences (lower left), and percentage measurement differences (lower right) for three comparisons of D50 measurements made as part of a QC assessment. Green and yellow bars represent means (n=5); black dots represent maximum and minimum values.

Conclusion: Literature sources point to high variability of this metric (Lanigan et al., 2010; Pleus and Schuett-Hames, 1998, Roper et al. 2010) and often recommend increasing the number of measured particles. The low variability within crew and between years in our data shows that the field protocol, specifically the sample size of 126 particles per sample reach and the use of the gravelometer, are appropriate. However, regardless of sample size, if the median particle (i.e., D_{50}) size is very close to the boundary between two size classes, then a single sample could change the D_{50} from one class to another, increasing the variability in the data. This doesn't entirely explain the 47% between-crew variability. No explanation is readily available.

Recommendation: Add other descriptors of coarse channel substrate, which are less prone to influences of particle size categorization and are more informative and relevant to fish habitat needs than D_{50}. Improve training to increase consistency between crews.

COARSE CHANNEL SUBSTRATE: PARTICLE SIZE DISTRIBUTION

- Averaged across all five sample reaches, particle size classifications were relatively consistent among surveys.
- The largest difference (5.5 percentage points) occurred between crews for the gravel (fine) class.
- The second-largest difference (4.6 percentage points) occurred between crews for the cobble class.

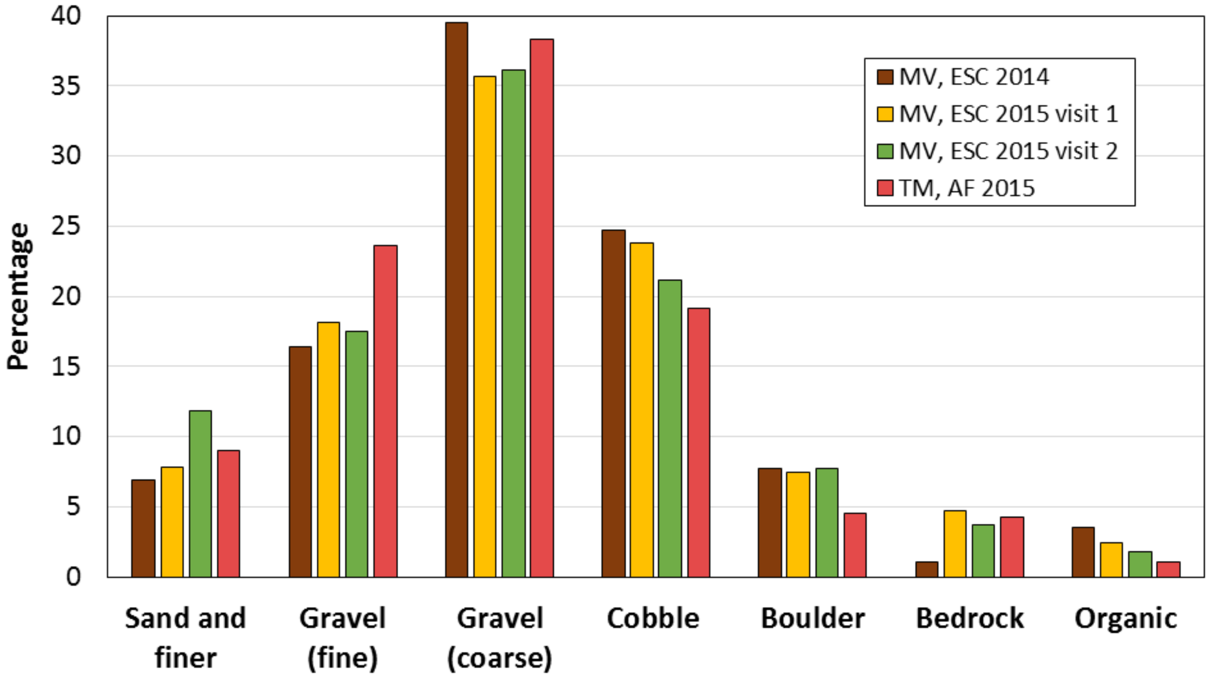

Figure 10. Percentage of substrate samples in each of seven particle size classes, including bedrock and organic materials, sampled during four visits as part of a QC assessment. Values represent the mean of five sample reaches.

Conclusion: Differences within the same year (2015) could be a result of crews removing samples from the cross-section and then not returning them. This would especially affect the cobble size class which is small enough to be removed but also large enough to gather smaller sediment around it. It could also explain why the cobble particle size class decreased between visits in 2015 while the gravel particle size class increased.

Recommendation: Reporting the changes in the coarse substrate by size class, as shown in Figure 10, may be a better way to track the quality of spawning habitat over time, than the D_{50} metric alone. If streams are sampled multiple times within the same year, replace particle samples during the sampling process, rather than depositing them downstream.

COARSE CHANNEL SUBSTRATE: PERCENT FINES

- The only significant comparison difference in percent fines occurred within crew and year. This was a result of consistently higher percent fines recorded in the second 2015 survey.
- The comparison differences in percent fines between crews and between years were not consistently in the same direction and were not significant.
- Large proportion of residual variance in the between-years comparison.

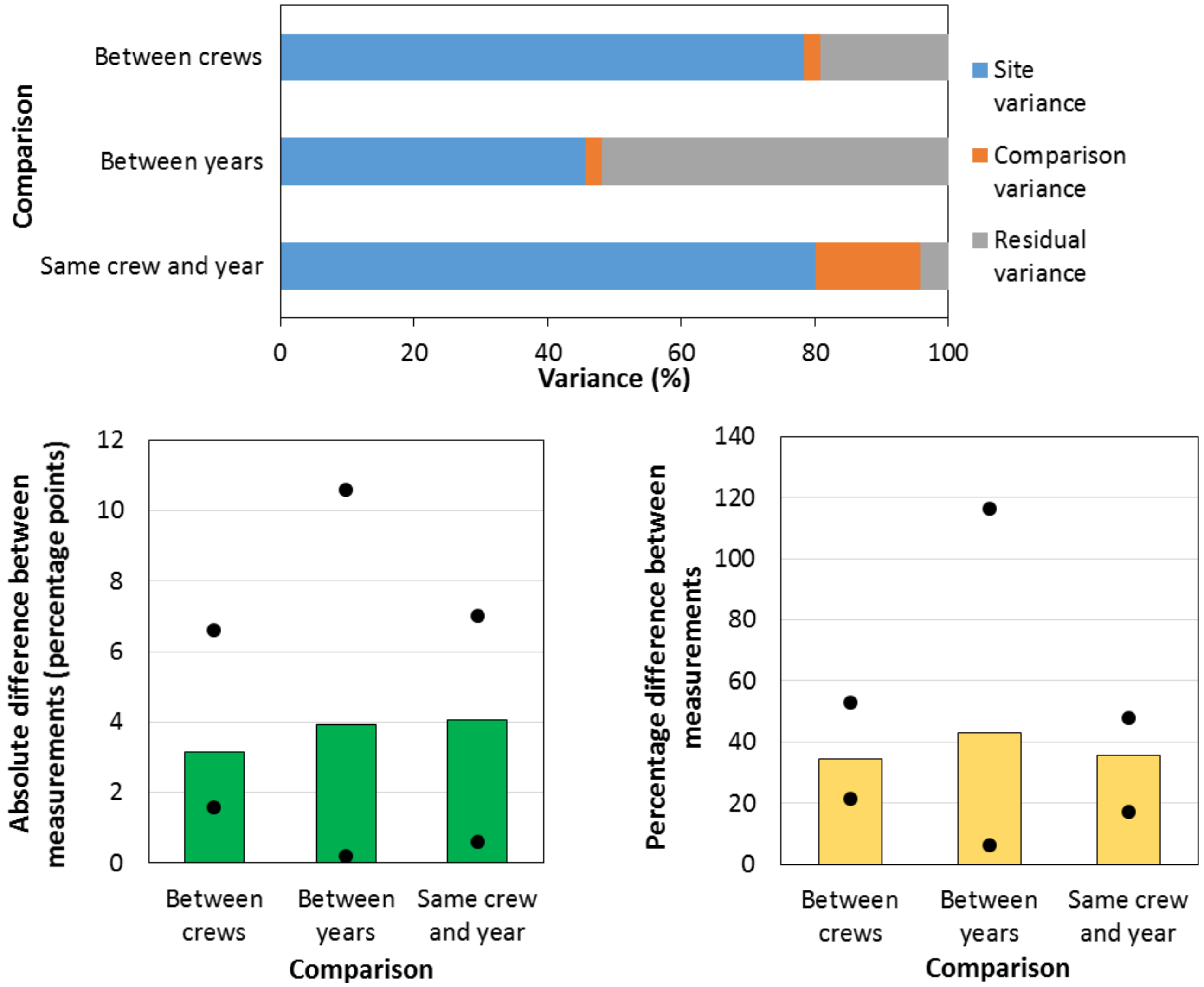

Figure 11. Variance distribution (top), measurement differences (lower left), and percentage measurement differences (lower right) for three comparisons of percent fines measured as part of a QC assessment. Green and yellow bars represent means (n=5); black dots represent maximum and minimum values.

Conclusion: The result that a significant difference in percent fines exists only in the same crew-same year comparison indicates high degree of random variation in this metric, likely due to the small sample size (only 7-12% of the 126 particles sampled per reach are fines). This also is the likely explanation for the large residual error in the between-years comparison.

Recommendation: Consider other field protocols for documenting fines. A large acceptable margin of error should be applied for this metric in the future trends analyses, especially if the field procedure is not changed.

COARSE CHANNEL SUBSTRATE: EMBEDDEDNESS

- Within the same crew and year, mean embeddedness differed by 3 to 10 percentage points among the three particle size classes for which embeddedness was recorded. The largest difference occurred for the boulder class.
- Between years, the largest embeddedness difference was 15 percentage points (boulder size class).
- Between crews, the difference in mean embeddedness ranged from 16 to 22 percentage points for the three particle size classes. The largest difference was for the boulder size class.

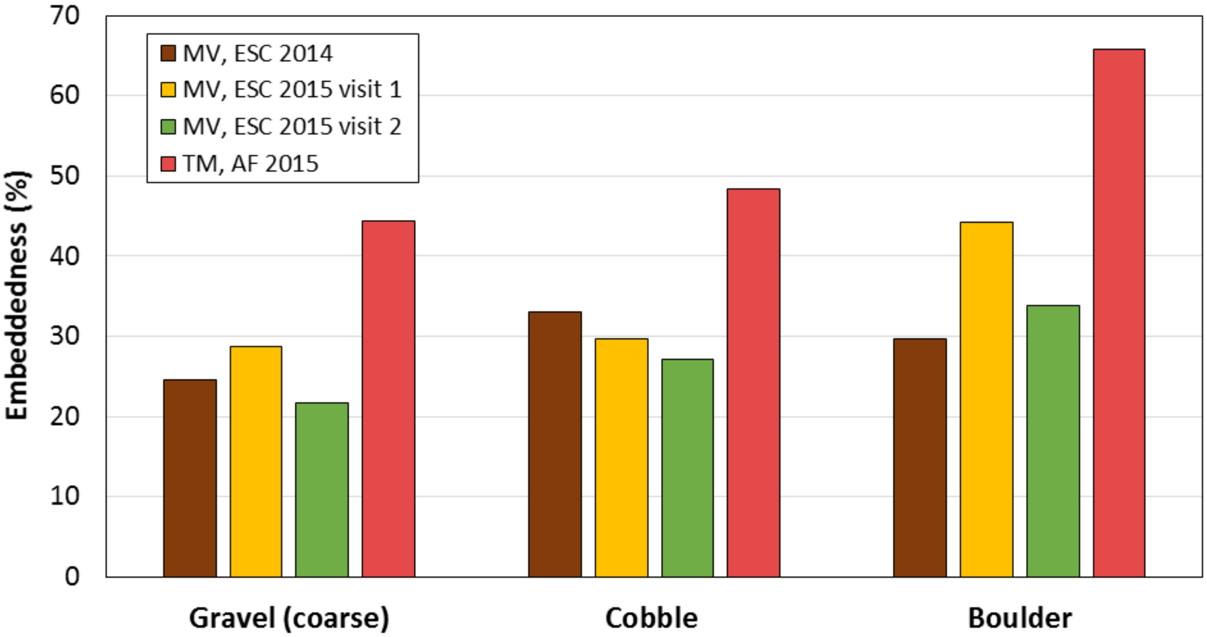

Figure 12. Percent embeddedness of substrate samples in each of three particle size classes, sampled during four visits as part of a QC assessment. Values represent the mean of five sample reaches.

Conclusion: This substrate characteristic is inherently variable because it is estimated by the observer. The differences are expected to be highest in the largest size class because many of the large boulders are buried and the estimation is even less precise. Still, the fact that inter-observer and between-year differences are considerably smaller than the between-crew ones, indicates that there is room for improvement of consistency by better training and better description of the field procedures.

Recommendation: Conduct comprehensive field training on this protocol if the field crew changes. Continue the annual calibration field training if the field crew is the same. Improve the description of the field procedure in the monitoring protocol. Consider only recording embeddedness on particles that are picked up, so that the demarcation line can be clearly observed; otherwise, values are guesses and may be biased.

IN-STREAM LARGE WOOD: TOTAL SINGLE PIECES/100 M (EXCLUDING PIECES IN JAMS)

- There was a significant difference in LWD piece counts between years; on average, seven more pieces per 100 m were measured in 2015 than in 2014. There was no significant difference in the number of pieces per 100 meters between crews or within the same crew and year.
- The percentage differences in piece counts for the three comparisons differed by an average of 11 to 23%.
- The greatest differences in piece counts all occurred in Basin 724 (represented by the three upper black dots on the graphs below). The greatest difference occurred between crews (a 24-piece difference).

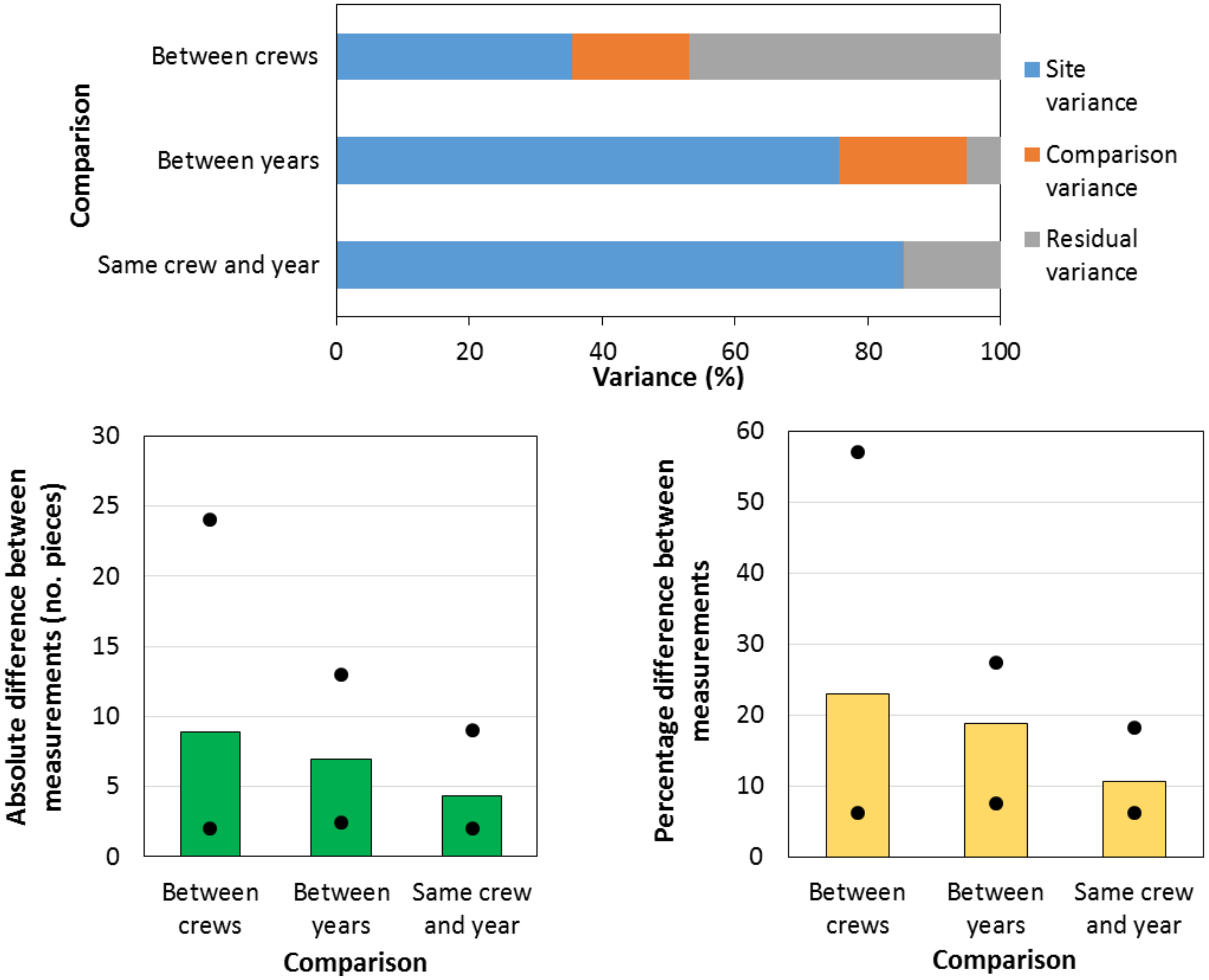

Figure 13. Variance distribution (top), measurement differences (lower left), and percentage measurement differences (lower right) for three comparisons of LWD pieces/100 m (excluding jams) measured as part of a QC assessment. Green and yellow bars represent means (n=5); black dots represent maximum and minimum values.

Conclusion: A large tree fell across one of the cross-sections in basin 724 between the 2014 and 2015 measurements. It altered the flow, which may have moved LWD pieces, and smashed/broken other LWD pieces in the vicinity. This may help explain the between-year difference of 19% .However, it cannot explain the between-crew variance. The different jam piece counts between crews (an 11-piece difference across the 5 basins) partially contributed.

Recommendation: Accept a relatively large margin of error for this metric in future trend analyses. Improve the description of qualifying pieces in the protocol. Improve field training.

IN-STREAM LARGE WOOD: TOTAL PIECES/100 M (INCLUDING PIECES IN JAMS)

- There was no significant difference in the number of pieces per 100 m (including pieces in jams) between crews or within the same crew and year.
- There was a significant difference in total pieces per 100 m between years; there was an average of 42 pieces per 100 m in 2014 and 48 pieces per 100 m in 2015.
- The percentage differences in piece counts for the three comparisons differed by an average of 10 to 23 percent. The greatest difference between crews was in basin 488 (28 pieces, a 50% difference); because this metric included pieces in jams, the difference in reported jams between crews (1 vs. 0) is unlikely to explain the piece count difference.

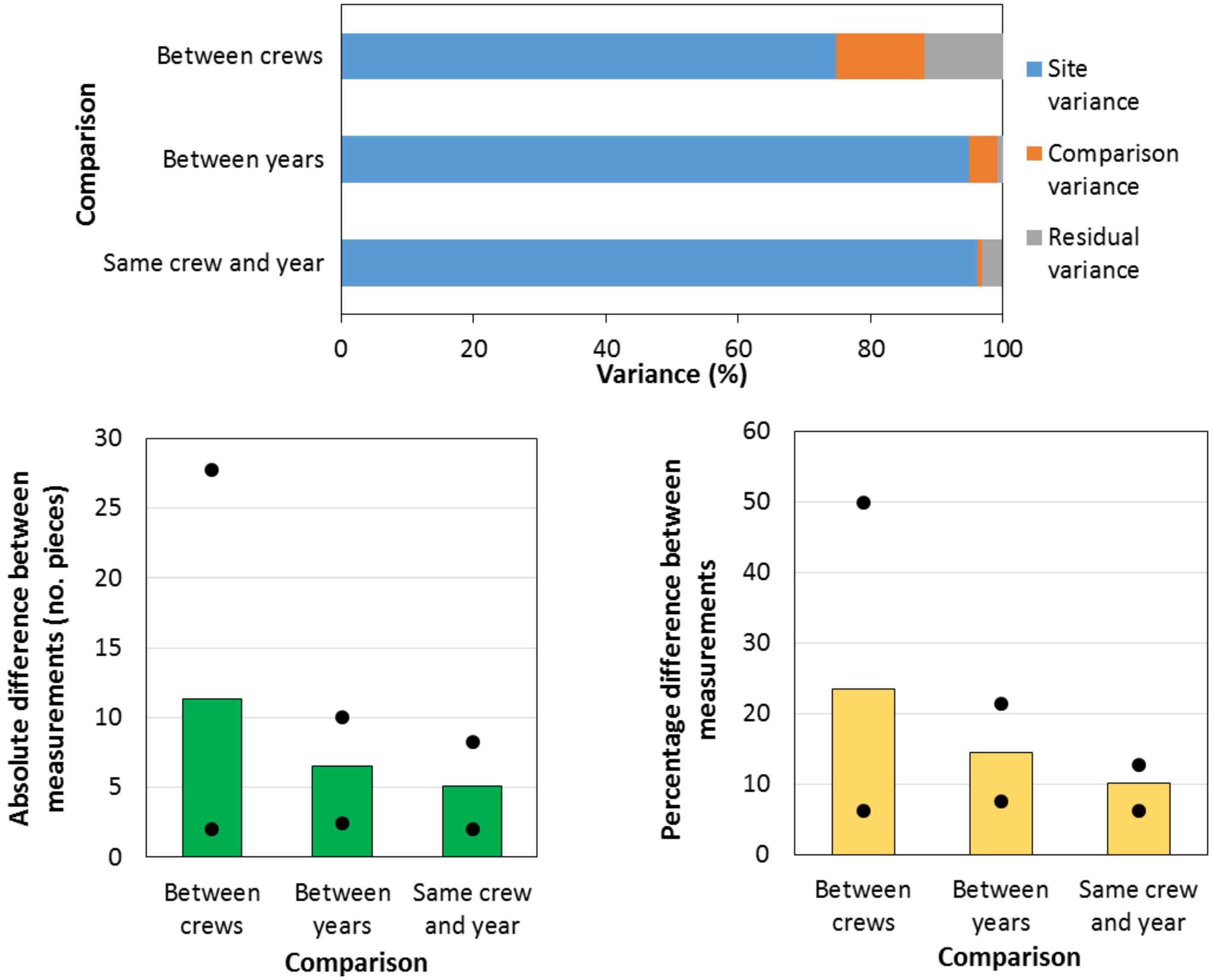

Figure 14. Variance distribution (top), measurement differences (lower left), and percentage measurement differences (lower right) for three comparisons of LWD pieces/100 m (including pieces in jams) measured in a QC assessment. Green and yellow bars represent means (n=5); black dots represent maximum and minimum values.

Conclusion: By including the number of pieces in jams (contrast with the previous metric above), the consistency in piece counts between crews was not improved.

Recommendation: Improve the description of qualifying pieces in the protocol. Improve field training.

IN-STREAM LARGE WOOD: NUMBER OF JAMS AND PIECES IN JAMS

- Two of the five sample reaches, 488 and 724, had jams. The number of jams in these two sample reaches differed among surveys (Table 3).
- In sample reach 488, the 2015 same crew-same year survey apparently recorded the same jam in both surveys, as it was recorded in segment A-B both times. The 2014 survey recorded two jams in 488, one in segment B-C and one in segment C-D. These two jams may have washed downstream into segment A-B between the 2014 and 2015 surveys.
- In sample reach 724, the surveys recorded either one or two jams. All of the jams recorded in this sample reach were in segment C-D.
- In sample reach 488, the total number of LWD pieces in jams ranged from 0 to 35 (Table 4).
- In sample reach 724, the number of pieces in jams was more consistent among surveys (13 to 24) than in 488. A least one jam in 724 had a piece count near the minimum value of 10 required to meet the definition of a jam; the fact that this was a "borderline" jam may have contributed to the different jam counts in this sample reach.
-

Table 3. Number of jams recorded per sample reach.

	Sample Reach				
Survey	**158**	**488**	**718**	**724**	**763**
MV, ESC 2014	0	2	0	1	0
MV, ESC 2015 visit 1	0	1	0	1	0
MV, ESC 2015 visit 2	0	1	0	1	0
TM, AF 2015	0	0	0	2	0

Table 4. Total number of LWD pieces recorded in all jams.

	Sample Reach					
Survey	**158**	**488**	**718**	**724**	**763**	**Total**
MV, ESC 2014	0	34	0	16	0	50
MV, ESC 2015 visit 1	0	35	0	13	0	48
MV, ESC 2015 visit 2	0	22	0	14	0	36
TM, AF 2015	0	0	0	24	0	24

Conclusion: Jams were not identified consistently.

Recommendation: Improve protocol and field training to consistently identify jams.

IN-STREAM LARGE WOOD: SPECIES CLASS

- Piece classification by species was relatively consistent between years and within the same year and crew, with differences of only 5 percentage points or less.
- Piece classification differed substantially between crews. Although the deciduous class differed by only a small amount (11% vs. 8%), the difference between crews for the coniferous class was large (24% vs. 60%), as was that of the unknown class (65% vs. 32%).

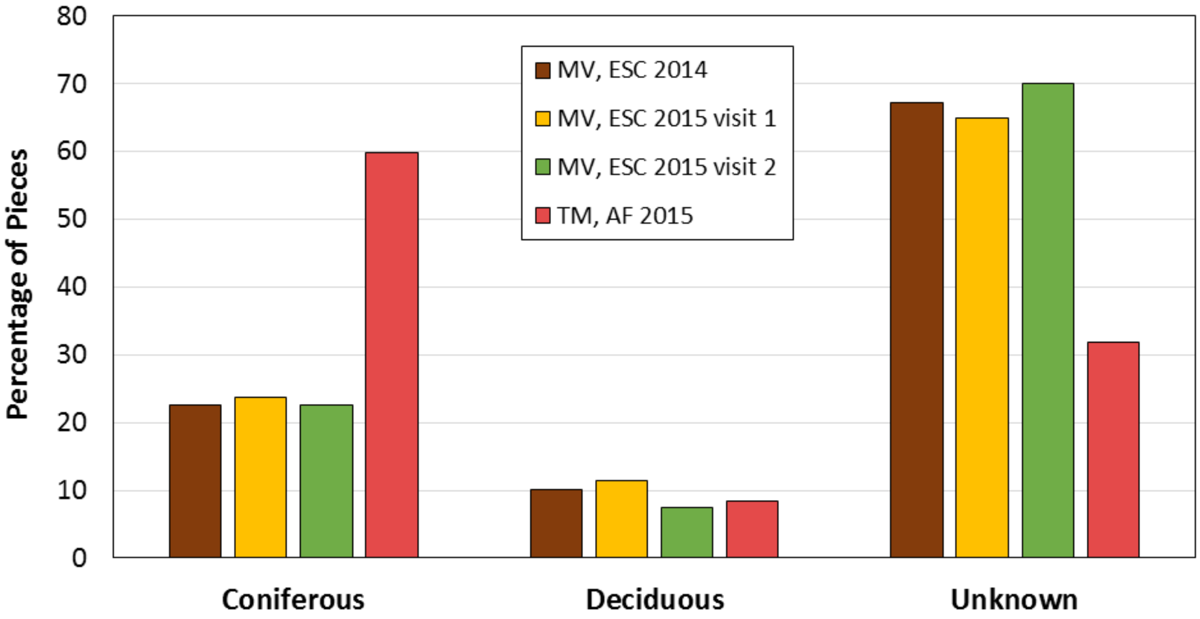

Figure 15. Percentage of LWD pieces in each of three species classes, sampled during four visits as part of a QC assessment. Values represent the mean of five sample reaches.

Conclusion: The TM, AF crew attributed pieces to the coniferous class more often than classifying them “unknown”, compared to the MV, ESC crew.

Recommendation: Improve field training to ensure the crew knows the classifying criteria and applies them consistently.

IN-STREAM LARGE WOOD: DECAY CLASS

- Piece classification by decay class was relatively consistent within decay classes 1, 2 and 3. Within each of those classes, none of the three comparisons yielded more than a four percentage point difference.
- For decay classes 4 and 5, differences were larger. In decay class 4, there was a 17 percentage point difference between crews and an 11 percentage point difference within crew and year. In decay class 5, there was a 13 percentage point difference between crews.

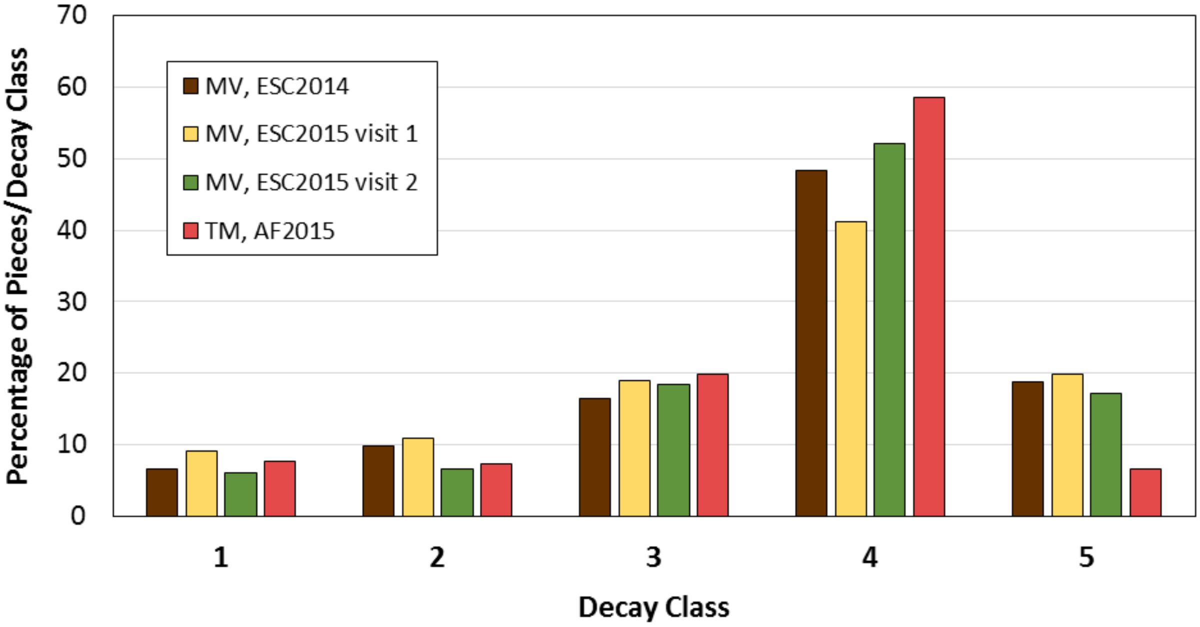

Figure 16. Percentage of LWD pieces in each of five decay classes, sampled during four visits as part of a QC assessment. Values represent the mean of five sample reaches.

Conclusion: The decay class is visually estimated and therefore the classification is inherently variable.

Recommendation: Improve field training to ensure the crew knows the classifying criteria.

IN-STREAM LARGE WOOD: MEAN PIECE DIAMETER

- There were no significant differences in mean piece diameter for any of the comparisons in the three models.
- Among the three comparisons, the largest average difference in mean piece diameter was 4.5 cm (13%), which occurred between crews. However, the measurement differences between crews were not uniformly in the same direction.

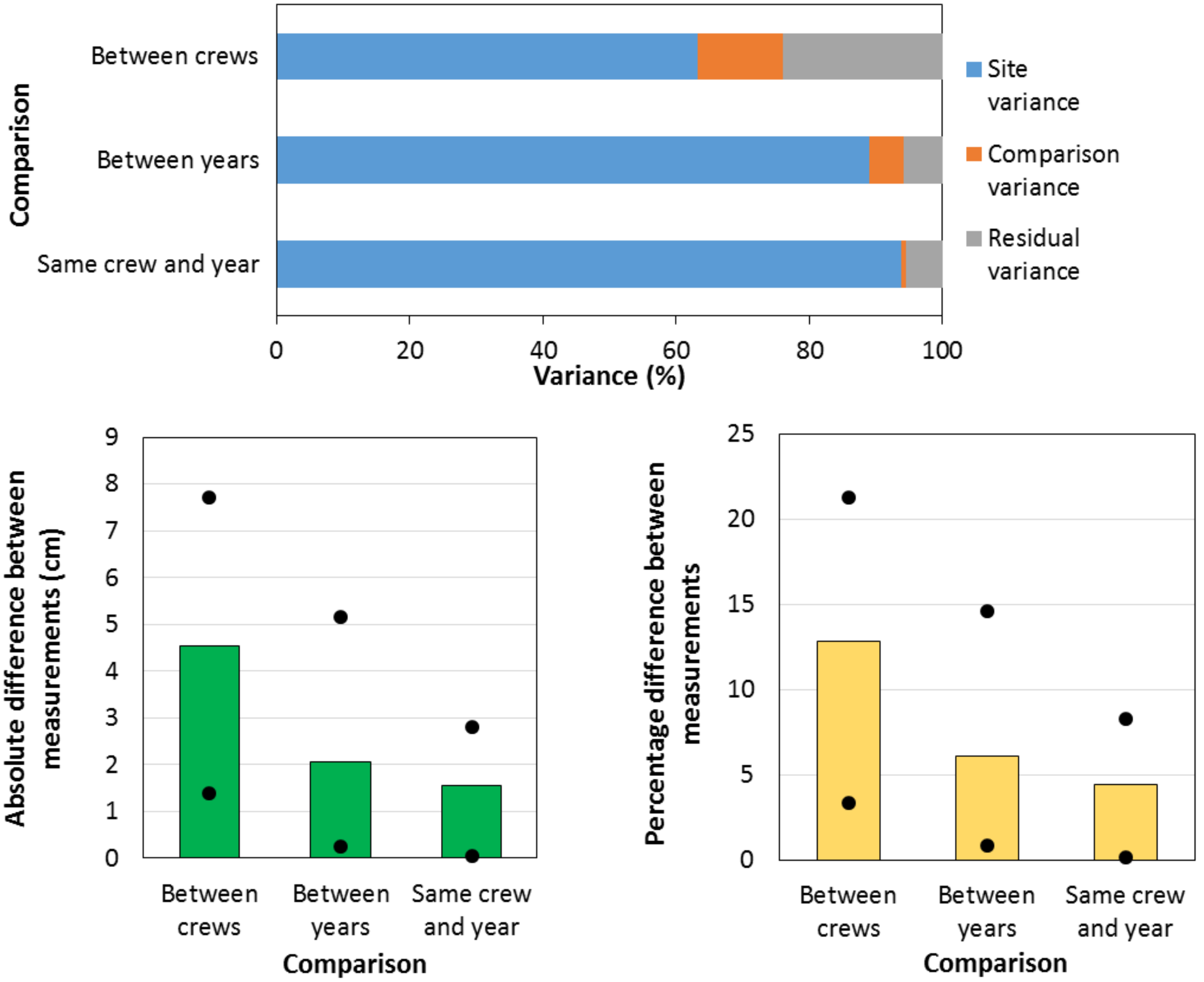

Figure 17. Variance distribution (top), measurement differences (lower left), and percentage measurement differences (lower right) for three comparisons of LWD mean piece diameter, measured in a QC assessment. Green and yellow bars represent means (n=5); black dots represent maximum and minimum values.

Conclusion: A likely reason for the difference between crews is where the diameter is measured (whether the piece midpoint is identified correctly) and whether the caliper is oriented consistently perpendicular to the piece central axis. This has an effect on measurements because most logs are not perfectly round.

Recommendation: Improve field training to increase the measurement precision. Describe potential sources of error in the field procedure section of the monitoring protocol.

IN-STREAM LARGE WOOD: ZONE 1 CUMULATIVE PIECE LENGTH/100 M

- There were no significant differences for zone 1 cumulative piece length for any of the comparisons in the three models.
- Percentage differences in each of the three comparisons were large (65% to 124%) but the differences were not consistently in the same direction and thus could not be well-explained by the model.
- Differences in cumulative piece length were influenced by different jam counts.

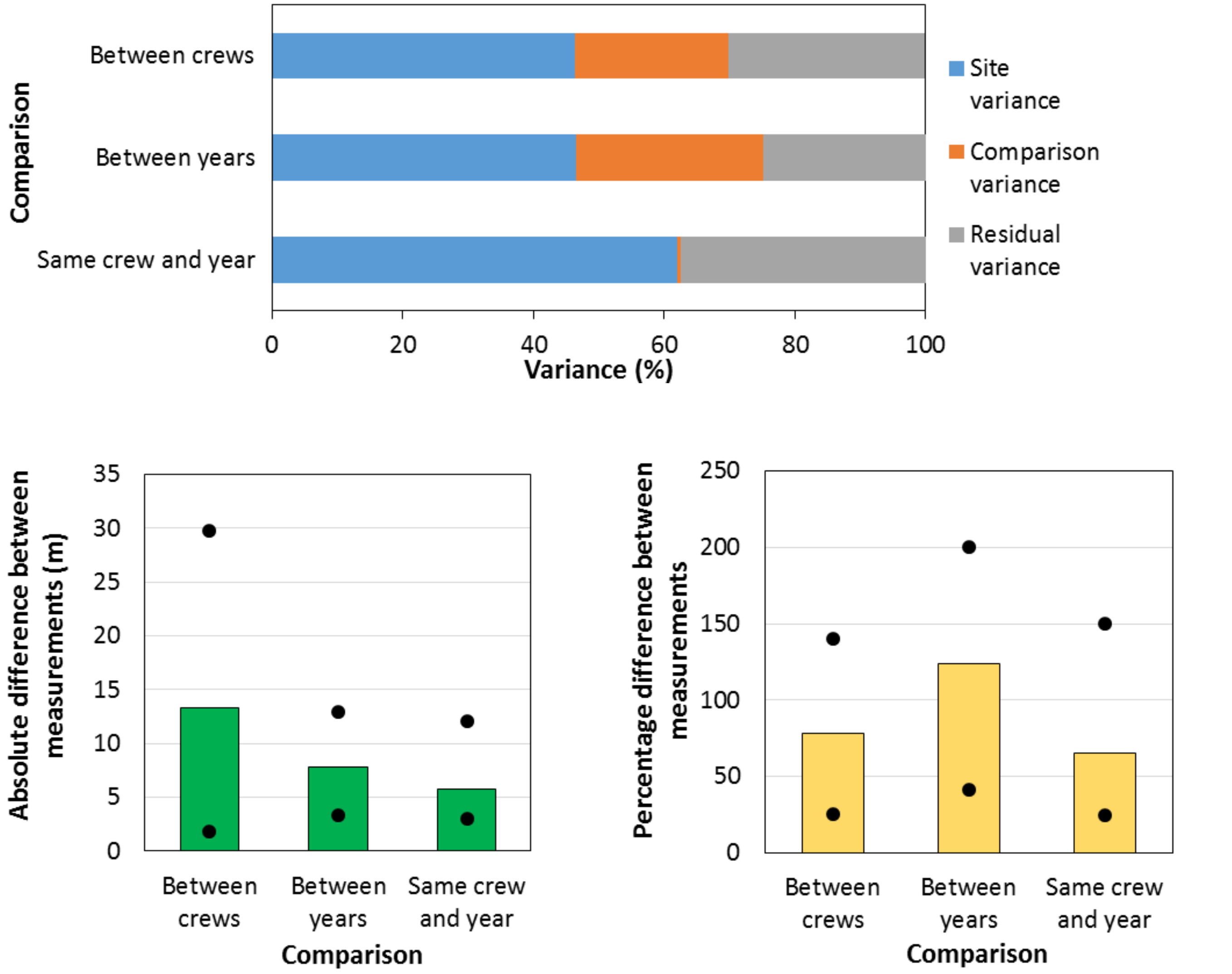

Figure 18. Variance distribution (top), measurement differences (lower left), and percentage measurement differences (lower right) for three comparisons of zone 1 cumulative LWD piece length, measured in a QC assessment. Green and yellow bars represent means (n=5); black dots represent maximum and minimum values.

Conclusion: The large differences between crews were partially due to differences in defining jams. The overall higher variability of this metric may be due to the small number of measurements taken in a single zone, which makes this metric very sensitive to one or two missed pieces. Additionally, measurement precision affected the results. One crew measured to the nearest 0.1 meter, whereas the other crew measured to the nearest 0.5 or 1 meter. Finally, identification of bankfull stage can affect zone 1 piece classification.

Recommendation: Improve field training, including training to consistently identify jams. Clearly define measurement precision level.

IN-STREAM LARGE WOOD: ZONE 2 CUMULATIVE PIECE LENGTH/100 M

- There was a significant difference between crews for zone 2 cumulative piece length per 100 m: the TM, AF crew measured, on average, 24 m more per sample reach than the MV, ESC crew. The largest difference (46 m; 113%) was in sample reach 488; this is likely because the TM, AF crew did not identify a log jam in the A/B segment, whereas the MV, ESC crew did.
- There was no significant difference between years or within the same crew and year, because the large differences were not consistently in the same direction. There was a substantial amount of residual variance in those models.
- Percentage differences in each of the three comparisons were large (56% to 68%). The *smallest* difference within the same crew and year was 41% (sample reach 718).

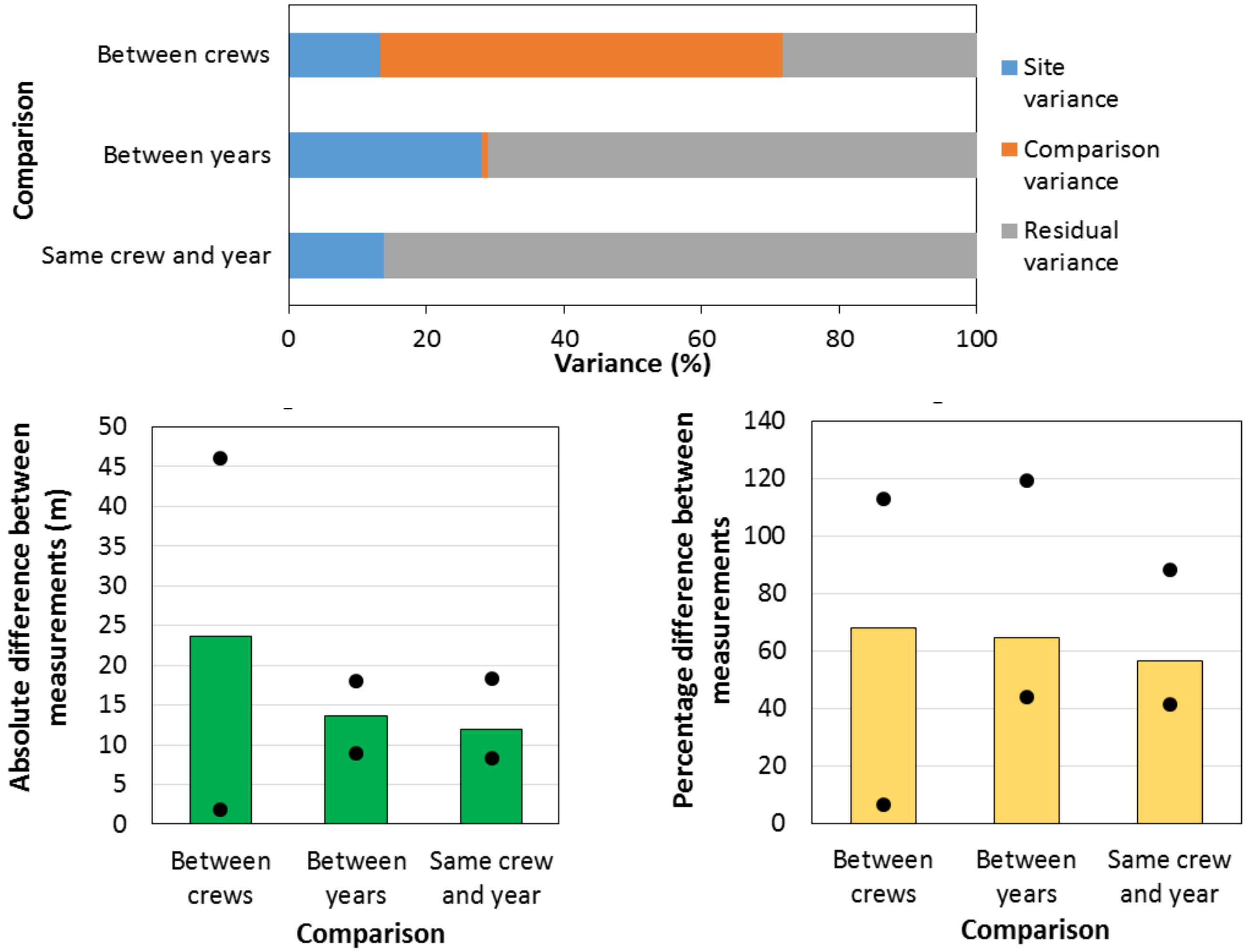

Figure 19. Variance distribution (top), measurement differences (lower left), and percentage measurement differences (lower right) for three comparisons of zone 2 cumulative LWD piece length, measured in a QC assessment. Green and yellow bars represent means (n=5); black dots represent maximum and minimum values.

Conclusion: Differences in defining jams had some effect on the between-crew differences, but does not explain the overall high level of variability in all three comparisons. It may be due to the small number of measurements taken in a single zone. The problem with measurement precision, described for the previous metric, may affect the results.

Recommendation: Improve field training. Consider modifying the protocol to facilitate better consistency. Clearly define measurement precision level.

IN-STREAM LARGE WOOD: ZONE 3 CUMULATIVE PIECE LENGTH/100 M

- There was a significant crew difference for zone 3 cumulative piece length per 100 m: the TM, AF crew measured, on average, 45 m *less* per sample reach than the MV, ESC crew. The largest difference (61 m; 91%) was in sample reach 158. This crew difference was the reverse of the pattern observed for zone 2 cumulative piece length.
- There was a relatively small, but statistically significant between-year difference, with an average of 8 m (14%) more per sample reach measured in 2015 than in 2014. There was no difference within the same crew and year. Overall the measurements difference were far greater between crews than between years or within the same crew and year.

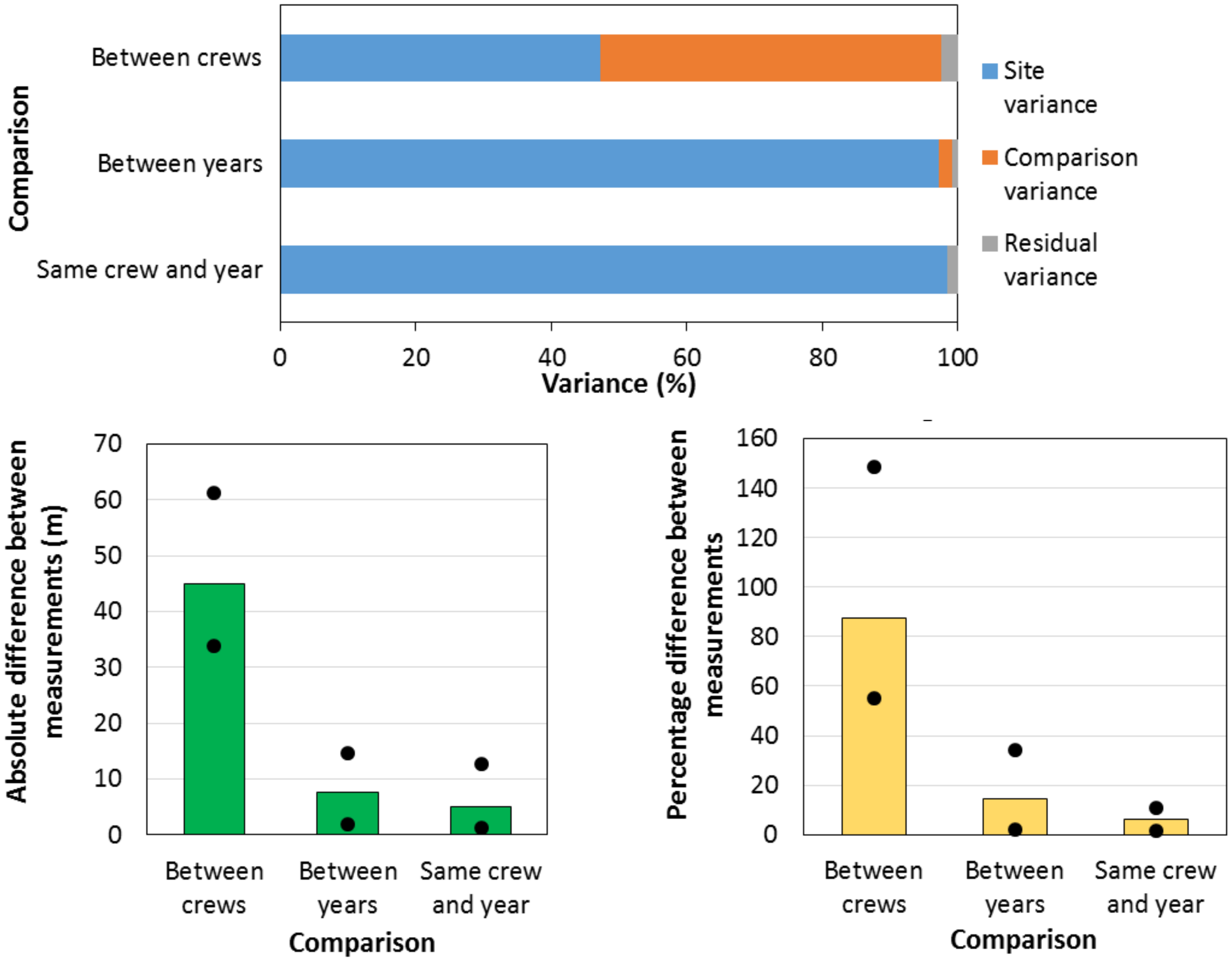

Figure 20. Variance distribution (top), measurement differences (lower left), and percentage measurement differences (lower right) for three comparisons of zone 3 cumulative LWD piece length, measured in a QC assessment. Green and yellow bars represent means (n=5); black dots represent maximum and minimum values.

Conclusion: There was a very large difference between crews that could be a result of the protocol itself or of the implementation of the protocol in the field. The overall higher variability of this metric may be due to the small number of measurements taken in a single zone and the inconsistent measurement precision level described above.

Recommendation: If this protocol is not modified to facilitate measurement consistency, then it will be important to improve field training. Clearly define measurement precision level.

IN-STREAM LARGE WOOD: ZONE 4 CUMULATIVE PIECE LENGTH/100 M

- There were no significant differences in zone 4 cumulative piece length for any of the comparisons in the three models.
- For the three comparisons, average differences in cumulative zone 4 piece length per 100 m ranged from 21 m (22%) within crew and year to 28 m (28%) between crews.

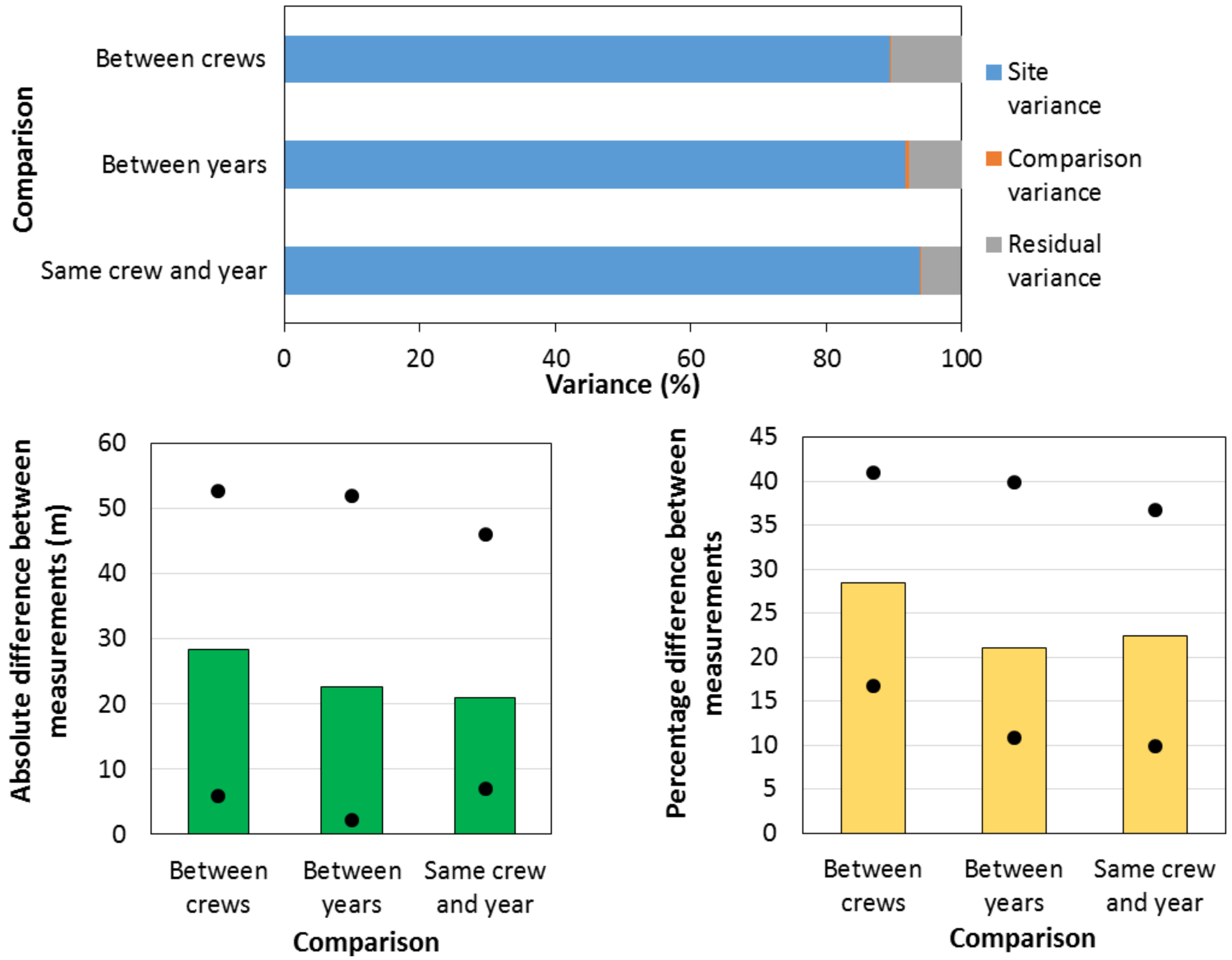

Figure 21. Variance distribution (top), measurement differences (lower left), and percentage measurement differences (lower right) for three comparisons of zone 4 cumulative LWD piece length, measured in a QC assessment. Green and yellow bars represent means (n=5); black dots represent maximum and minimum values.

Conclusion: Differences in zone 4 were smaller (expressed as a percentage) than the differences in zones 1, 2 or 3. The cumulative length of LWD in zone 4 was much greater, on average (mean = 99 m), than in zones 1 and 2 (mean = 11 and 28 m, respectively), therefore making zone 4 cumulative length less sensitive to one or two missed pieces.

Recommendation: Clarify protocol and improve field training. Clarify measurement precision level.

IN-STREAM LARGE WOOD: POOL-FORMING PIECES/100 M

- There were no significant differences in the counts of pool-forming pieces per 100 m for any of the comparisons in the three models.
- For the three comparisons, the average differences in the count of pool-forming pieces per 100 m ranged from 2.1 to 2.6 (53% to 91%).

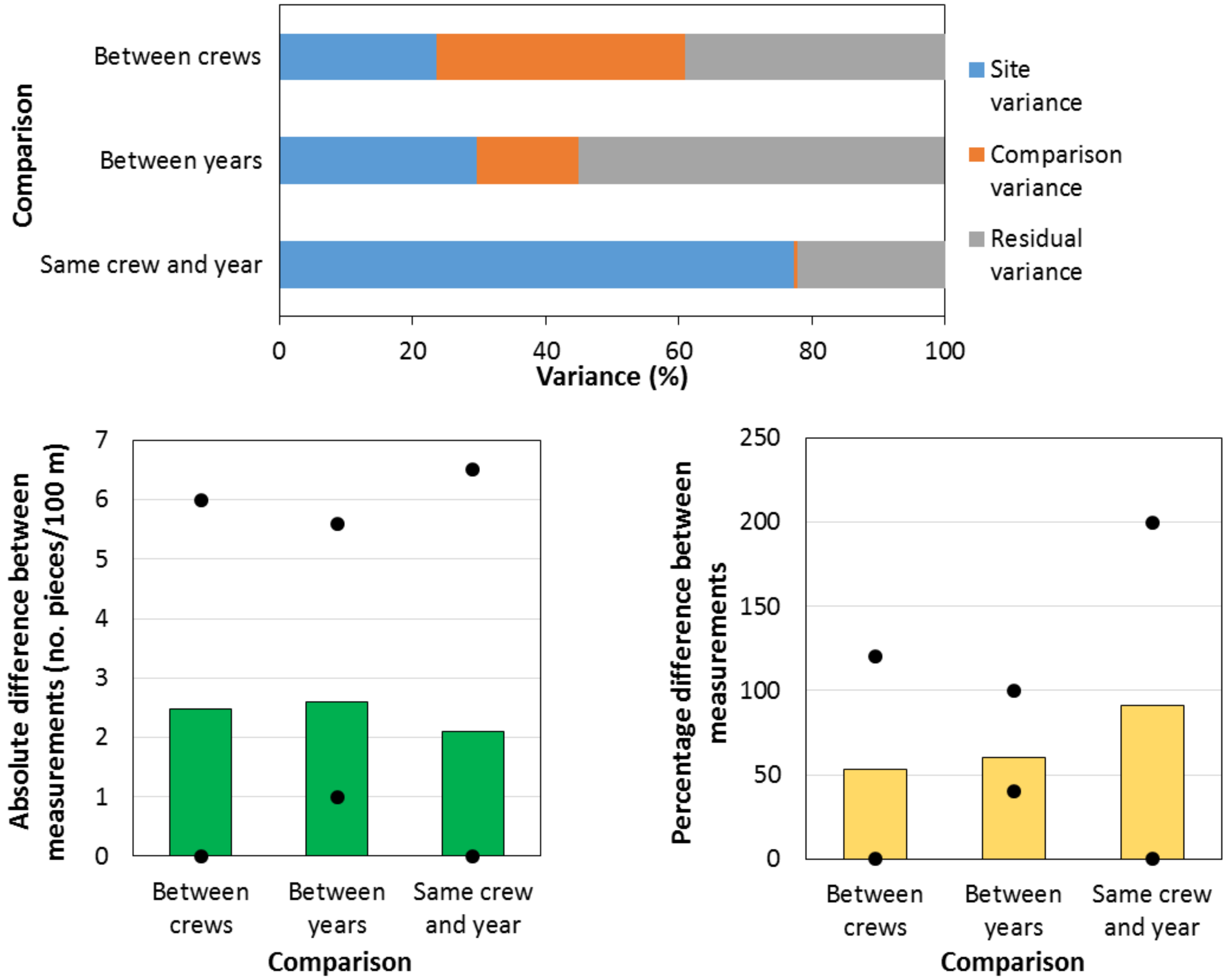

Figure 22. Variance distribution (top), measurement differences (lower left), and percentage measurement differences (lower right) for three comparisons of the count of pool-forming LWD pieces/100 m, measured in a QC assessment. Green and yellow bars represent means (n=5); black dots represent maximum and minimum values.

Conclusion: The typically small number of pool-forming pieces per 100 m of sample reach (less than 4, on average) means that a difference of only 2 or 3 pieces in a survey can lead to a very large percentage difference. This also is the likely explanation for the large residual errors in all 3 comparisons.

Recommendation Consider a large margin of error for this metric in the future analyses. Consider dropping this metric because it is not repeatable under the current protocol.

IN-STREAM LARGE WOOD: SEDIMENT-STORING PIECES/100 M

- There were no significant differences in the count of sediment-storing pieces per 100 m for any of the comparisons in the three models.
- For the three comparisons, the average differences in the count of sediment-storing pieces per 100 m ranged from 2.6 (24%) between years to 6.0 (43%) between crews.

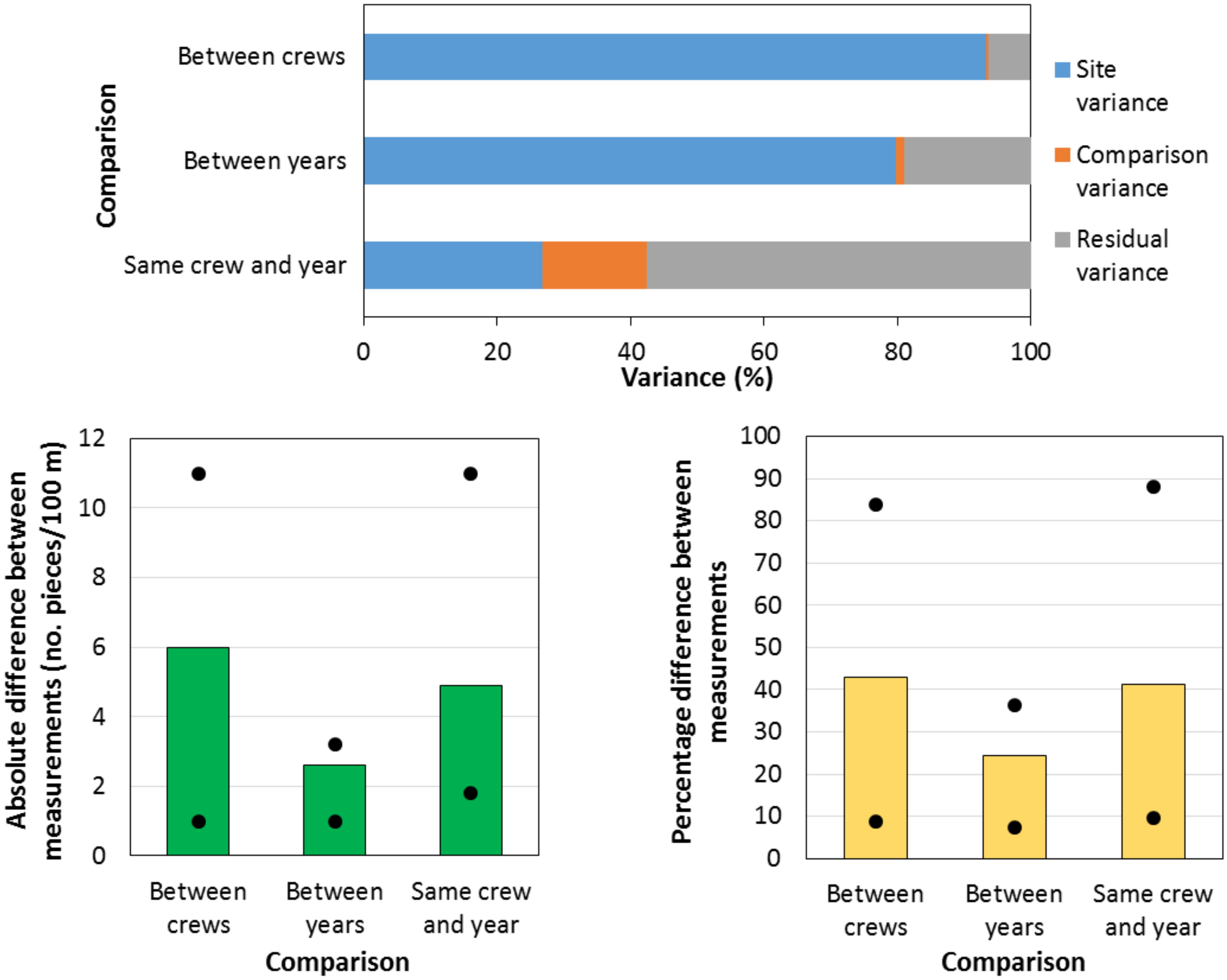

Figure 23. Variance distribution (top), measurement differences (lower left), and percentage measurement differences (lower right) for three comparisons of the count of sediment-storing LWD pieces/100 m, measured in a QC assessment. Green and yellow bars represent means (n=5); black dots represent maximum and minimum values.

Conclusion: The relatively small number of sediment storing pieces per 100 m of sample reach (mean of 13.4 pieces) may contribute to the high variability, because misidentification of only a few pieces could significantly affect the result. Improved field training may increase the accuracy in classifying a piece.

Recommendation Consider a large margin of error for this metric in the future analyses. Improve field training to increase classification accuracy.

IN-STREAM LARGE WOOD: ORIENTATION OF PIECES

- Within all five orientation classes, the largest comparison difference was consistently the between-crew difference. Averaged across all five classes, the between-crew difference was 10.2 percentage points. The TM, AF crew averaged a larger percentage of pieces in classes B and C, and a smaller percentage of pieces in the A, D, and vertical classes, compared with the MV, ESC crew.
- Between years, the average difference across all classes was 2.7 percentage points. Within crew and year, the average difference across all classes was 2.2 percentage points.

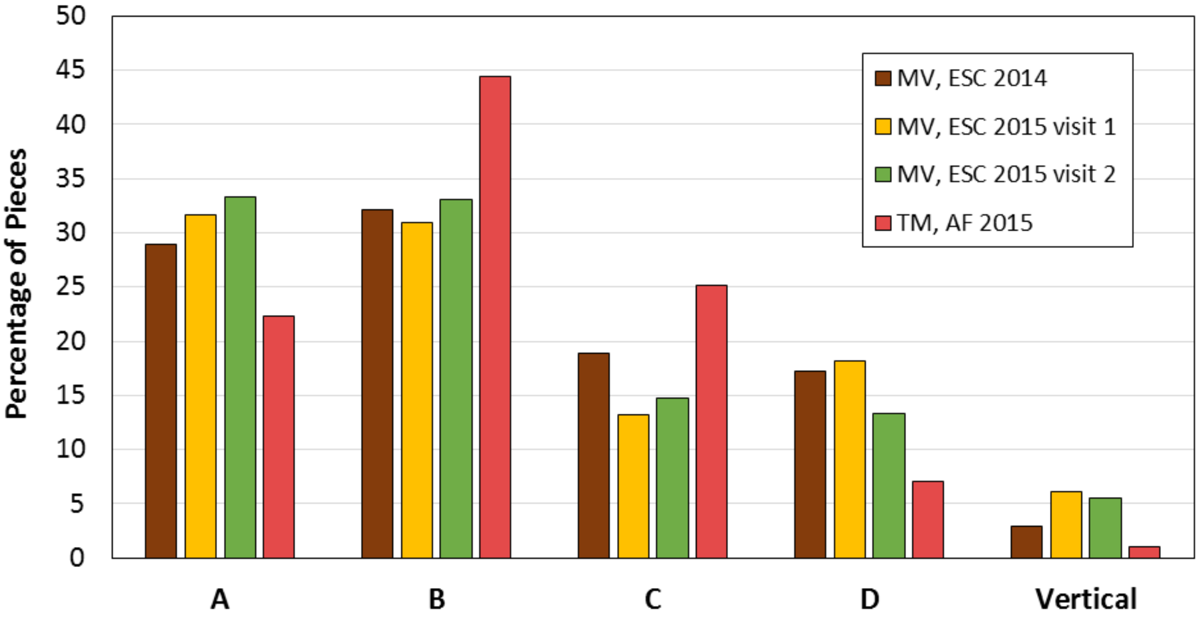

Figure 24. Percentage of LWD pieces in each of five orientation classes, sampled during four visits as part of a QC assessment. Values represent the mean of five sample reaches.

Conclusion: Piece orientation is visually estimated and therefore the classification is inherently variable. Yet, the large between-crews difference, compared to the ones within crew and between years, indicates that the classification accuracy can be improved through training.

Recommendation: Improve field training to increase the classification accuracy.

LWD: STABILITY OF PIECES

- Across the five stability classes, the largest comparison difference was the between-crew difference, which averaged 9.9 percentage points. The between-crew difference ranged from 1.2 percentage points (rooted class) to 21.9 percentage points (stable class).
- Between years, the average difference across all five classes was 5.7 percentage points. The maximum between-year difference was 13.1 percentage points in the stable class.
- Within crew and year, the average difference across all five classes was 2.0 percentage points.

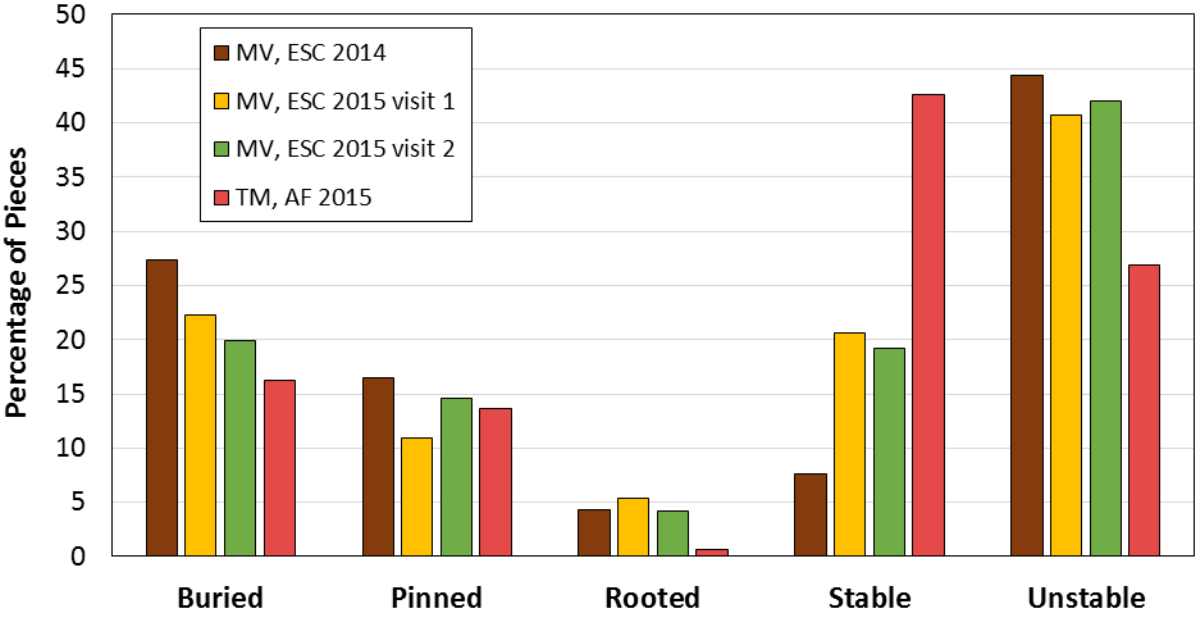

Figure 25. Percentage of LWD pieces in each of five stability classes, sampled during four visits as part of a QC assessment. Values represent the mean of five sample reaches.

Conclusion: The stability class is subjective and therefore the classification is inherently variable. The TM, AF crew classified more pieces as stable pieces and fewer as unstable. Mitchell Vorwerk noted: “Buried, pinned, rooted are all fairly stable logs. ‘Stable’ is just the more general term or the term for large logs that aren’t going anywhere that don’t fit into the other stable classifications. In addition to visual, we usually give the log a ‘kick test’ to help determine how stable it is.”

Recommendation: Improve field training to increase the classification accuracy. In the protocol, define better the category “stable” to exclude the other cases. All categories have to be mutually exclusive.

HABITAT UNITS: NUMBER OF HABITAT UNITS PER 100 M

- This metric was based on the total count of habitat units per sample reach. It was calculated by dividing the total length of all habitat units by the total count of habitat units, then adjusting the result to a 100-m basis.
- There were no significant differences in count of habitat units/100 m for any comparison in the three models.
- Among the three comparisons, the largest difference was between crews, at 4.4 habitat units per 100 m (a 33% difference). The TM, AF crew averaged 12.2 habitat units per 100 m, whereas the MV, ESC crew averaged 14.6 units per 100 m. However, the TM, AF crew did not always count fewer units; in sample reach 763, they counted 16 units per 100 m and the MV, ESC crew counted 11.1 units per 100 m.
- The largest difference between crews was in basin 488: the MV, ESC crew counted 17.2 units per 100 m, and the TM, AF crew counted 10.6 units per 100 m.
- The largest difference in unit count within the same crew and year occurred in basin 763: counts were 15.7 units per 100 m and 11.1 units per 100 m.

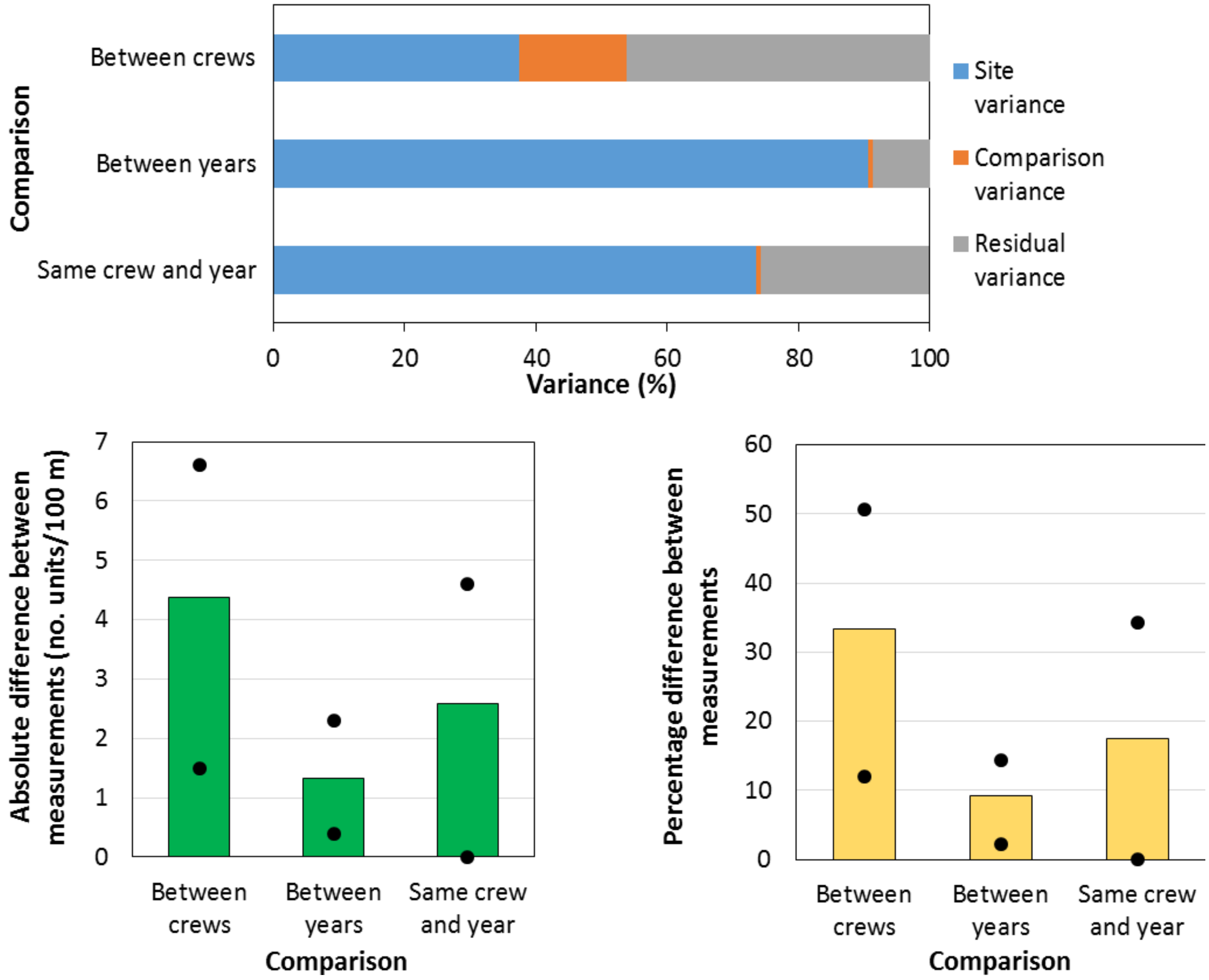

Figure 26. Variance distribution (top), measurement differences (lower left), and percentage measurement differences (lower right) for three comparisons of the count of habitat units/100 m, measured in a QC assessment. Green and yellow bars represent means (n=5); black dots represent maximum and minimum values.

Conclusion: This metric is inherently variable since the habitat units' identification is based on qualitative criteria. Other monitoring studies also report low consistency of this metric (Roper et al. 2010) and specifically, the correct identification of a unit and the identification of starting and ending points (Archer et al 2004). The differences between our crews is likely due to lumping vs. splitting habitat units. It takes only one lumping of several habitat units into one long unit to introduce a big difference between crews. Different flow levels can also have a substantial effect on habitat unit identification.

Recommendation: Repeat the sampling in each basin at approximately the same time of the year to minimize the seasonal effect of stream flow. Improve field training, including more training materials and time for training. Focus training on the lumping vs. splitting issue. Make sure that one member of the crew is designated to make the final call on the unit type. Consider a large acceptable margin of error for this metric.

HABITAT UNITS: NUMBER OF POOLS PER 100 M

- This metric was based on the total count of pools (scour pools, dammed pools, and backwater pools) per sample reach. The count was then adjusted to a 100-m basis.
- There were no significant differences in count of pools/100 m for any comparison in the three models.
- Among the three comparisons, the largest mean difference was between years, at 1.7 pools per 100 m (a 31% difference). The second-largest difference was between crews, at 1.6 pools per 100 m (29%).
- The largest difference between crews was in basin 763: the TM, AF crew counted 7.0 pools per 100 m, and the MV, ESC crew counted 3.7 pools per 100 m. The largest difference in pool count within the same crew and year also occurred in basin 763: counts were 5.5 pools per 100 m and 3.7 pools per 100 m.

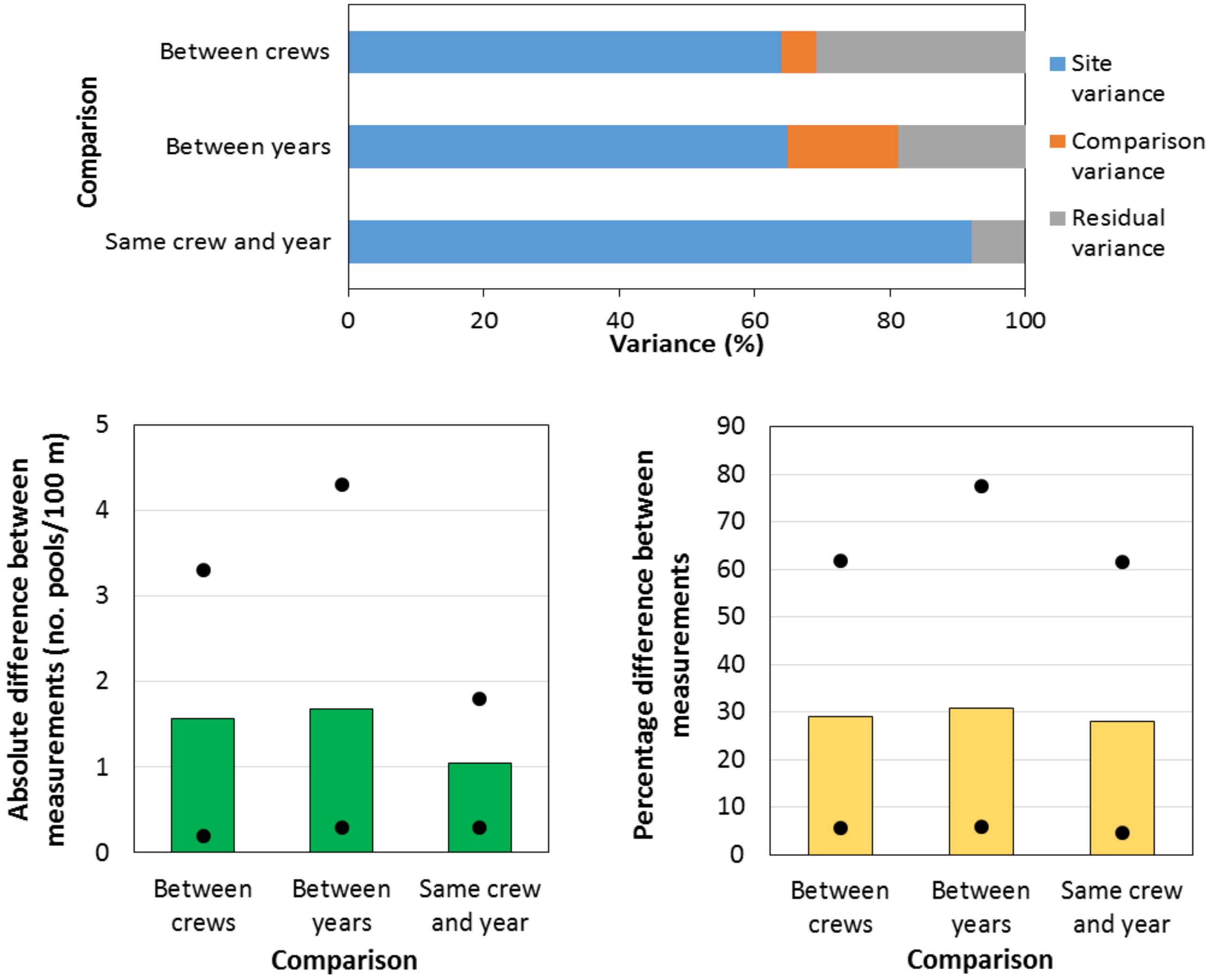

Figure 27. Variance distribution (top), measurement differences (lower left), and percentage measurement differences (lower right) for three comparisons of the count of pools/100 m, measured in a QC assessment. Green and yellow bars represent means (n=5); black dots represent maximum and minimum values.

Conclusion: Different flow levels can also have a substantial effect on this metric which may explain the between year difference and the difference between crews (the first crew sampled in mostly low flow conditions and the second crew sampled at higher water levels).

Recommendation: Repeat the sampling in each basin at approximately the same time of the year to minimize the seasonal effect of stream flow. Improve field training, including more training materials and time for training. Focus training on the lumping vs. splitting issue. Make sure that one member of the crew is designated to make the final call on the unit type.

HABITAT UNITS: NUMBER OF HABITAT UNITS PER 100 M BY TYPE

- The largest difference for any of the three comparisons occurred within the scour pool type: in 2015, an average of 2.0 fewer scour pool (SP) units per 100 m were recorded, compared with 2014.
- The next largest comparison difference occurred within the run (RN) unit type. The MV, ESC crew recorded an average of 1.6 more run units per 100 m, compared with the TM, AF crew.
- In 2015, an average of 1.4 more run units per 100 m were recorded, compared with 2014.

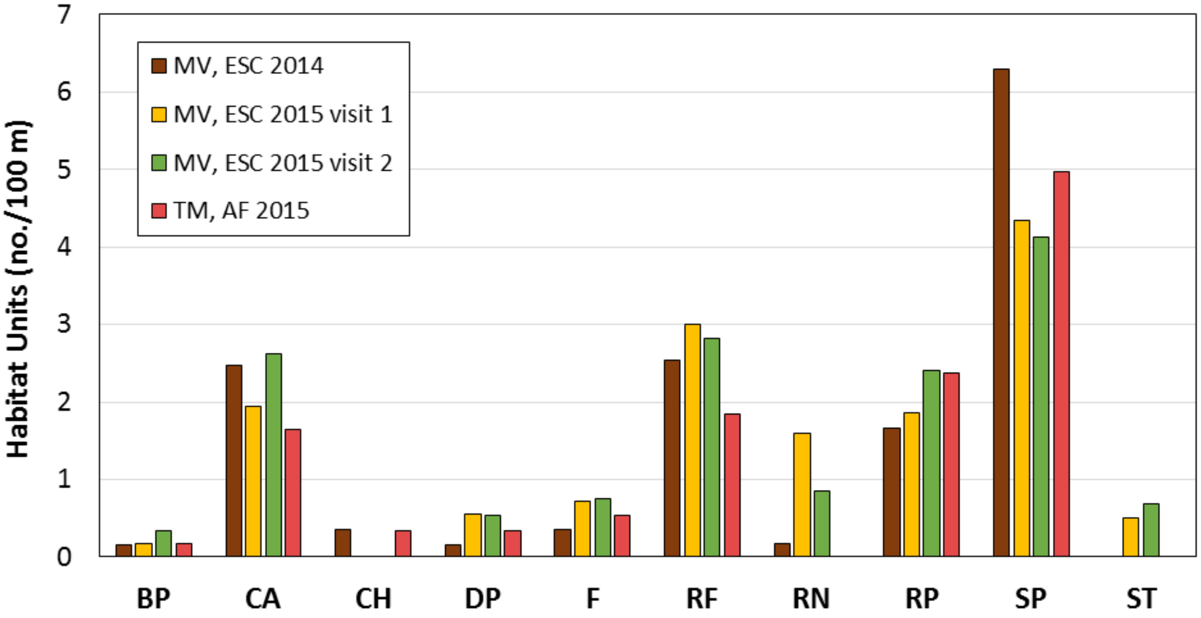

Figure 28. Number of habitat units per 100 m in each of ten classes, sampled during four visits as part of a QC assessment. Values represent the mean of five sample reaches.

Conclusion: This metric is inherently variable since the habitat units' identification is based on qualitative criteria. Lower flow in 2015 compared to 2014 may partially explain the smaller number of identified pools. The larger between-crews difference compared to the ones within crew and between years, indicates that the classification accuracy can be improved through training.

Recommendation: Repeat the sampling in each basin at approximately the same time of the year to minimize the seasonal effect of stream flow. Improve field training including more training materials and time for training. Focus training on the lumping vs. splitting issue. Make sure that one member of the crew is designated to make the final call on the unit type. Consider a large acceptable margin of error for this metric.

HABITAT UNITS: PERCENTAGE OF SAMPLE REACH ALLOCATED TO EACH UNIT TYPE

- The largest difference for any of the comparisons occurred within the rapid (RP) type: the allocation of the sample reach to rapids was an average of 12 percentage points greater for the TM, AF crew, compared with the MV, ESC crew.
- Other large differences between crews occurred for riffles (RF) and runs (RN). On average, the MV, ESC crew recorded 24 percent riffles, whereas the TM, AF crew recorded 14 percent riffles. The MV, ESC crew recorded 10 percent runs, whereas the TM, AF crew recorded 0 percent runs.
- Within in the cascade (CA) type, there was a difference of 10 percentage points between surveys by the same crew in the same year. These differences in percent CA occurred in basin 763 (0% vs. 32% on visits 1 and 2, respectively) and in basin 724 (19% vs. 33% on visits 1 and 2, respectively). In both of these basins, the reductions in CA were associated with increases in RN and RP.
- There were two large differences between years. In 2014 scour pools (SP) averaged 38 percent of the sample reach, whereas in 2015 they averaged 26 percent. Runs averaged 1 percent of the sample reach in 2014 and 10 percent in 2015. This may be due to water level differences between years. In lower water, a run may be lumped with a rifle or pool.

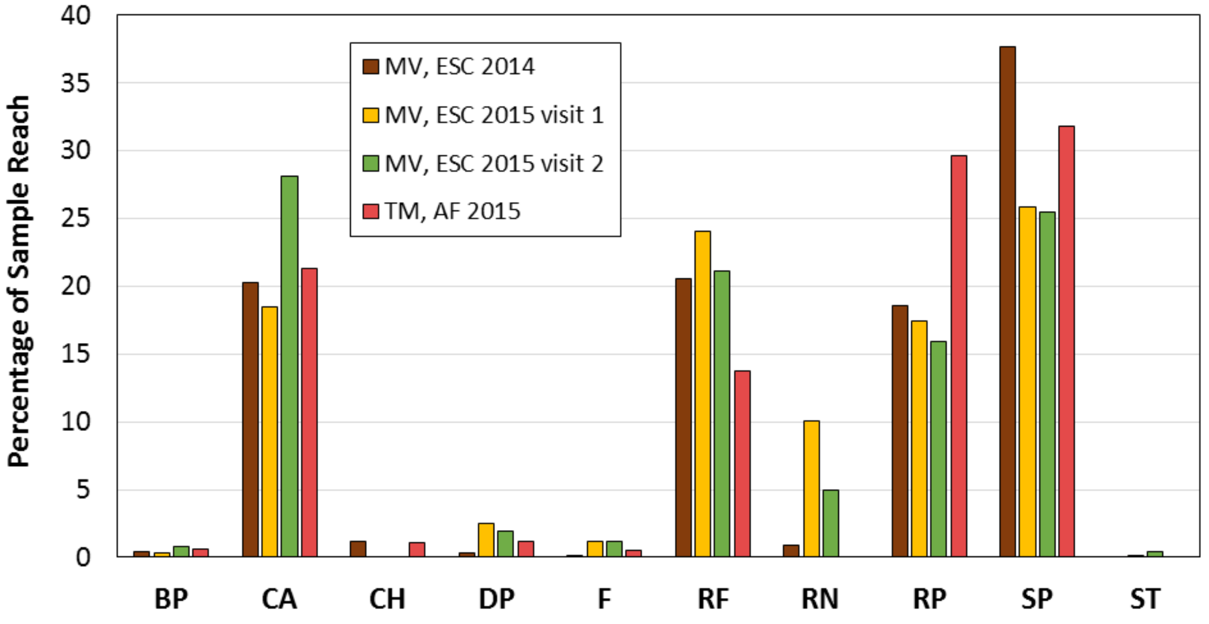

Figure 29. Percentage of sample reach allocated to each of ten habitat unit classes, sampled during four visits as part of a QC assessment. Values represent the mean of five sample reaches.

Conclusion: The TM, AF crew was classifying more units as rapids and fewer units as riffles and runs.

Recommendation: Improve field training including training materials and time for training. Make sure that one member of the crew is designated to make the final call on the unit type. Use caution using this metric for trends analysis since it is highly-prone to inter-observer error.

HABITAT UNITS: RESIDUAL POOL DEPTH

- There was no significant difference in residual pool depth for any of the comparisons in the three models.
- The greatest difference in residual pool depth measurements occurred between years: an average difference of 9.6 cm (26%). The largest difference between years was 24.2 cm (58%) in sample reach 763. This large difference was not associated with any obvious error in the data.
- Differences between crews and within the same crew and year were smaller, averaging less than 6 cm (15%).

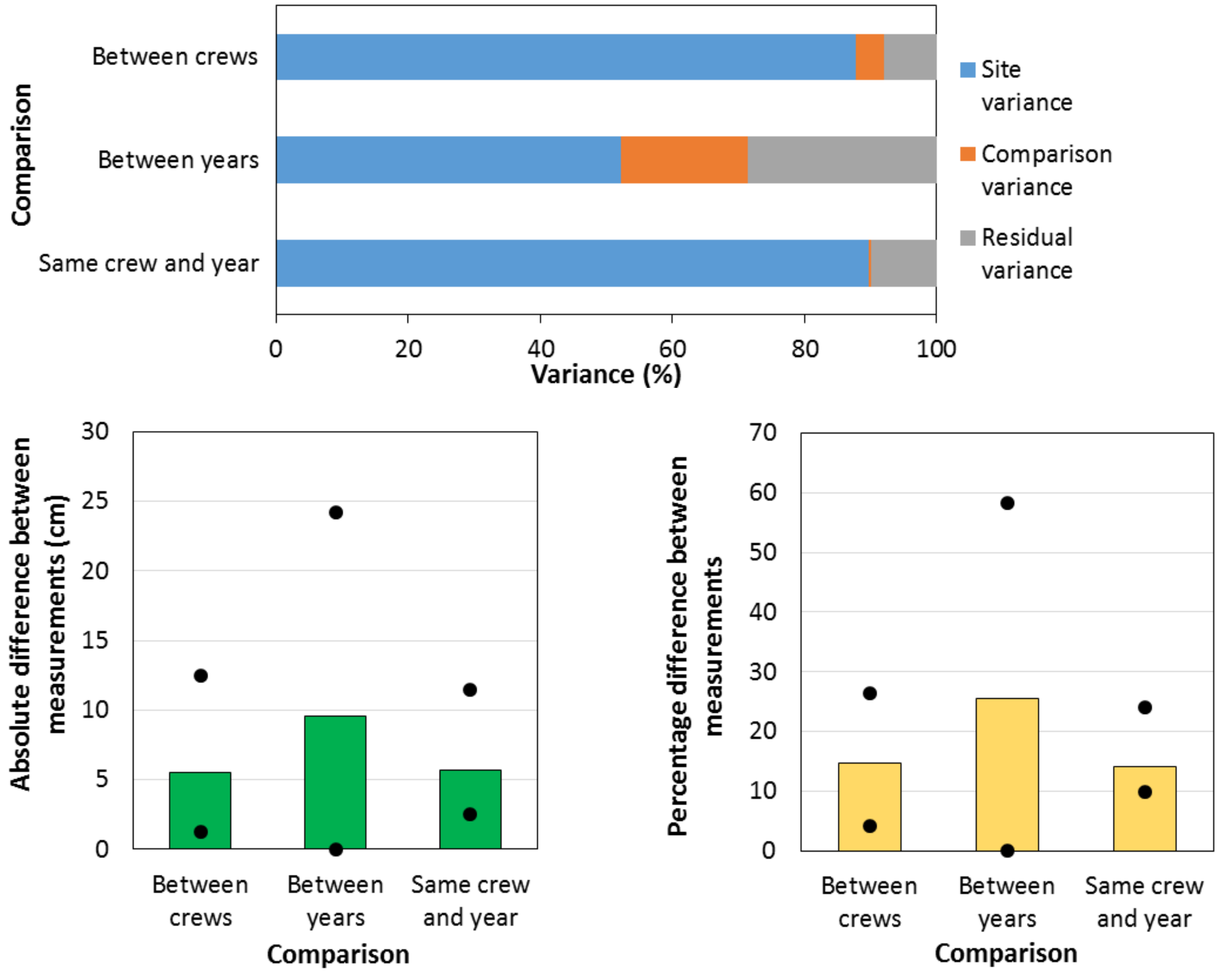

Figure 30. Variance distribution (top), measurement differences (lower left), and percentage measurement differences (lower right) for three comparisons of residual pool depth, measured in a QC assessment. Green and yellow bars represent means (n=5); black dots represent maximum and minimum values.

Conclusion: There was good consistency between crews and within crews. The higher between-year difference is expected because of the dynamic nature of the small streams geomorphology in the OESF. Different flow levels can also have effect on this metric. Part of the differences in all three comparisons may be explained by random variation due to the small sample size (4-6 pools per reach).

Recommendation: Repeat the sampling in each basin at approximately the same time of the year to minimize the seasonal effect of stream flow.

VALLEY SEGMENT AND CHANNEL TYPE CLASSIFICATION

- Valley segment type was uniformly recorded as Alluvial for all of the sample reaches.
- The channel type of each sample reach was classified uniformly among surveys, with the exception of sample reach 158, which differed between crews (Table 5). The different classifications in this sample reach were likely a result of the different gradient measurements between crews. The MV, ESC crew measured slope values ranging from 8.01% to 8.06%, whereas the TM, AF crew measured slope as 7.44%.

Table 5. Channel type classifications.

	Sample Reach				
Survey	**158**	**488**	**718**	**724**	**763**
MV, ESC 2014	Cascade	Pool-riffle	Pool-riffle	Step-pool	Step-pool
MV, ESC 2015 visit 1	Cascade	Pool-riffle	Pool-riffle	Step-pool	Step-pool
MV, ESC 2015 visit 2	Cascade	Pool-riffle	Pool-riffle	Step-pool	Step-pool
TM, AF	Step-pool	Pool-riffle	Pool-riffle	Step-pool	Step-pool

Conclusion: High consistency in the channel type categorization is important because the categories are used to infer other characteristics of the stream (such as hydraulic power) and to inform about the nature of the ecological processes (such as sediment transport). The overall consistency is good but may be improved further.

Recommendation: Improve the field training including materials and time for training to ensure accurate classification of channel types.

RIPARIAN VEGETATION (OVERSTORY)

- The initial 2015 measurement failed to tag and measure between 2.5% and 6.7% of trees per plot (Table 6). This calculation assumes that none of the tags fell off or were removed from the trees between the initial measurement and the QC check; this is a possibility because, in basin 718, some of the nails were discovered to be loose during the QC check.
- Differences in the number of trees per hectare between original measurements and QC checks were relatively small (<5%) on all plots except for those in basin 763, where differences were 14.8% and 29.8% (Table 2). This site was exceptionally difficult to measure, owing to the steep slopes on both plots and to the fact that one plot (cross-section A) was installed in a bend in the sample reach, so that the plot intersected the main stem of the river. Also, for safety reasons, the QC crew chose not to try to measure some trees on the other plot (cross-section C) which were located on an unstable slope.
- Differences in basal area per hectare between original measurements and QC measurements were approximately 10% or less on all plots except for those in basin 763, where the differences were 24.4% and 26.3% (Table 6). These differences are again attributed to the very steep slopes and the fact that the QC crew did not visit all of the trees for safety reasons.
- The 10.6% basal area difference in basin 584 was due primarily to differences in the DBH measurement of a single, very large tree (218.5 cm vs. 244.0 cm).
- Species identification was relatively consistent between the initial measurement and the QC check in basins 584 and 718 (Table 7). In basin 158, vine maple was misidentified as bigleaf maple in the initial measurement. In basins 158 and 763, there were clearly some instances of species misidentification among conifers during the initial measurement. Notably, Pacific silver fir was not identified at all in basin 763 during the initial measurement.

Table 6. Number of trees per basin from the summer 2015 crew measurement, and the number of trees missed during that measurement, as determined by the 2015 quality control check.

Basin	Trees recorded in 2015 measurement (no.)	Trees missed in original 2015 measurement (no.)	Percentage missed
158	155	4	2.5
584	145	4	2.7
718	182	13	6.7
763	210	10	4.5

Table 7. Summary of live trees on two 0.18-ha plots in each of four monitoring basins, measured by a summer 2015 crew and by a DNR QC crew. Any difference between the two crews' values is considered measurement error by the summer crew.

	Plot		Trees/ha			Basal area (m^2/ha)		
Basin	(cross-section)	Measurement	Result	Absolute error	% error	Result	Absolute error	% error
158	B	QC check	344			62.9		
		2015 measmnt.	339	-5	-1.6	67.3	+4.4	+6.8
	E	QC check	528			47.1		
		2015 measmnt.	511	-17	-3.2	45.1	-2.0	-4.3
584	B	QC check	400			63.3		
		2015 measmnt.	389	-11	-2.8	56.9	-6.4	-10.6
	F	QC check	394			73.9		
		2015 measmnt.	383	-11	-2.9	70.3	-3.6	-5.0
718	B	QC check	250			55.7		
		2015 measmnt.	250	0	0.0	58.5	+2.8	+4.9
	D	QC check	750			22.7		
		2015 measmnt.	722	-28	-3.8	22.2	-0.5	-2.2
763	A	QC check	444			41.0		
		2015 measmnt.	600	+156	+29.8	53.4	+12.4	+26.3
	C	QC check	383			29.1		
		2015 measmnt.	444	+61	+14.8	37.2	+8.1	+24.4

Table 8. Percentage of live trees, by species, in four basins measured by a summer 2015 crew and quality checked by a DNR QC crew.

Basin	Species	QC check	2015 measurement	Difference
		- - - - - - - - - - - - - - - % - - - - - - - - - - - - - - -		
158	ACCI	3.1	0	-3.1
	ACMA	0	3.2	+3.2
	ALRU	45.9	45.9	0
	PISI	8.1	3.2	-4.9
	PSME	4.4	1.3	-3.1
	TSHE	38.4	46.5	+8.1
584	ABAM	1.4	0.7	-0.7
	ABGR	0	0.7	+0.7
	ALRU	18.2	19.6	+1.4
	PISI	1.4	2.1	+0.7
	THPL	15.5	15.4	-0.1
	TSHE	63.5	61.5	-2.0
718	ALRU	1.1	1.1	0
	PISI	9.8	10.6	+0.8
	TSHE	89.1	88.3	-0.8
763	ABAM	11.9	0	-11.9
	ALRU	18.8	18.8	0
	PISI	18.8	36.9	+18.1
	PSME	13.1	3.4	-9.7
	THPL	3.8	3.4	-0.4
	TSHE	33.8	37.6	+3.8

Conclusion: It was apparent the plot installation protocol was not strictly followed in some cases. Three different crews were used in the 2015 field season. Two crews were relatively new at plot installation and were not continually checked by the project supervisor. Crew supervisors had a tendency to "improve" efficiency by shortcutting tagging protocols. Two possible sources of error were identified that threaten documenting stand dynamics: trees that were missed when the plot was installed and repeated diameter measure of large trees on slopes. Errors associated with tree tags that are lost between measurements are usually easily diagnosed and corrected. Errors in species determinations were minor and easily updated over time. Errors associated with repeated tree diameter measurement were very small and will likely remain so with subsequent re-measurements.

Recommendation: It would be valuable to have paid consistent crews to establish and re-measure plots. Careful initial tagging of trees and marking of the height measurement locations on large trees will improve stand change detection. Tag loss could be reduced by ensuring nails are securely within the wood, angled 10 degrees downward with the tag against the nail head. Re-measurement would be aided by tagging with sequential tags; dividing the plot into tagging lanes at the center line will help ensure individual trees are not missed. A fixed measurement height on very large trees would significantly reduce re-measurement error. Protocols for repositioning plots should be expanded, for both safety and non-safety reasons. For example, the right bank plot in basin 763 should have been relocated for safety considerations, and for the fact that the plot intersected the main stem river that ran perpendicular to the sample reach.

Discussion and general recommendations

Signal-to-noise ratios

All continuous stream survey metrics are compared by calculating S:N ratios for same-crew-and-year and for between-crew comparisons (Table 9). This analysis helps identify which need improved protocols and/or field training. The same-crew-and-year S:N represents the precision of a metric under ideal conditions, whereas the between-crew S:N represents the precision of a metric under more typical conditions of different crews interpreting and implementing the monitoring protocol. Roper et al. (2010) suggested that for S:N below 2.5, the metric has a low likelihood of detecting real differences in the condition of habitat attribute. Metrics with S:N between 2.5 and 6.5 are considered moderately consistent and those with S:N above 6.5 – highly consistent.

Table 9. Signal: noise ratios (S:N) for metrics based on continuous variables (metrics based on distribution among categories, such as substrate particle size class distribution, are not included). Metrics with low S:N values (<2.5) are shown in bold type.

Metric	**Same crew and year**	**Between crews**
Stream Morphology		
Channel gradient	40,318.1	53.1
Bankfull width	65.8	17.7
Bankfull depth	9.1	2.7
Floodplain width	6.7	2.5
Bankfull thalweg depth	18.3	3.6
Bankfull width: depth ratio	3.5	2.8
Erosion	10.2	4.6
Channel Coarse Substrate		
D$_{50}$ (median particle size)	76.8	2.7
Percent fines	4.0	3.6
In-Stream Large Wood		
Total pieces/100 m (excluding pieces in jams)	5.8	**0.5**
Total pieces/100 m (including pieces in jams)	24.8	3.0
Piece mean diameter	15.0	**1.7**
Total length of all pieces, per 100 m, in Zone 1	**1.6**	**0.9**
Total length of all pieces, per 100 m, in Zone 2	**0.2**	**0.2**
Total length of all pieces, per 100 m, in Zone 3	63.0	**0.9**
Total length of all pieces, per 100 m, in Zone 4	15.2	8.4
Pool-forming pieces per 100 m	3.4	**0.3**
Sediment-storing pieces per 100 m	**0.4**	14.0
Habitat Units		
Units per 100 m	2.8	**1.0**
Pools per 100 m	11.5	**1.8**
Residual pool depth	8.8	7.1

Only 3 metrics show low constituency between same crews (LWD total length in zones 1 and 2) and the number of sediment storing LWD pieces. The small sample size is a likely contributor and brings into question the sampling and analyses of LWD pieces by zone. Consider revising the protocol.

Stream Morphology

Channel gradient and bankfull width were highly consistent between crews (S:N > 6.5) and the rest stream morphology metrics were moderately consistent between crews (S:N between 2.5 and 6.5). Fine-tuning of the protocol and improved field training would likely benefit the metrics with S:N in the moderate range.

Recommendations: The criteria and training for identifying bankfull stage and floodplain width should be refined.

In-Stream Large Wood

Overall, the results for between-crew comparison of in-stream large wood indicate that protocol revision and improved field training is needed for all in-stream large wood metrics. It is unclear why the S:N is so high for a single metric (total piece length in Zone 4). This could easily be a random occurrence associated with our limited sample size, and therefore we do not recommend exempting this metric from protocol revision.

Recommendations: Review the need for each of the LWD metrics in light of what will be used in trends analysis and in future riparian validation monitoring (fish response to managed watersheds). Reconsider collecting and analyzing LWD data by zone. Consider removing the sediment storing function of LWD from the protocol or describe the classification criteria better to improve consistency.

Channel Coarse Substrate

Between-crew comparisons show moderate consistency of these metrics.

Recommendations: Review literature for additional metrics that can be calculated using data collected under the current protocol. Look for metrics that are informative about the quality of spawning habitat and sensitive to its changes over time. At a minimum, improve protocol and training to increase consistency in classifying embeddedness.

Habitat Units

The two habitat unit metrics that are dependent upon unit identification (unit count per 100 m and pools per 100 m) had low S:N, which also suggests protocol revision and improved field training is needed. Residual pool depth had a high S:N, suggesting that metric is relatively precise under the current protocol.

Recommendations: In the monitoring protocol, reference available field guides for classifying habitat units. Improve field training to increase consistency in identifying the type and dimensions of all units. Focus training on the issue lumping vs. splitting.

We compared our S:N values to those for the same or similar metrics, presented in Roper et al. (2010) for the Aquatic and Riparian Effectiveness Monitoring Program (AREMP), Environmental Monitoring and Assessment (EMAP), and Northwest Indian Fisheries Commission (NIFC) monitoring groups. There was a total of eight comparable metrics; for seven of these metrics, we compared our S:N to those from AREMP and EMAP. For the eighth (in-stream large wood piece count), we also compared our S:N to that of NIFC. For the following metrics, our S:N (between-crew comparison) values were equal to or greater than those of both AREMP and EMAP: bankfull width, bankfull width: depth ratio, D_{50}, percent fines, pools per 100 m, and residual pool depth. For one metric, channel gradient, our S:N was greater than that of EMAP's but smaller than that of AREMP's (although all three were exceptionally high). For the eighth metric, in-stream large wood piece count per 100 m including jam pieces, our S:N of 3.0 was smaller than that of NIFC (44.1) and AREMP (9.9) but greater than that of EMAP (0.9). Among the eight metrics in this comparison, we can conclude that the greatest emphasis for protocol revision and training improvement should be on in-stream large wood protocols.

Sources of variation

During the course of the field work and data analysis, we identified potential sources of variation (Table 10) that can be easily avoided or their influence can be reduced in the future field sampling through proper timing of the surveys, training of the field crews, and clarifications in the monitoring protocols.

Table 10. Potential sources of variation in each of the three QC comparisons.

Comparison	**Potential reason for measurement differences**
Between-crew variation (MV, ESC vs. TM, AF)	**Disturbances** from multiple crew visits (kicking substrate, removing substrate after measuring, eroding banks with foot traffic, breaking soft LWD pieces, etc.)
	Weather (attitude, speed of work) and different water heights/flow (LWD zones 1 and 2 fluctuating, different LEW heights, etc.)
	Electronic field form (Access file on Panasonic ToughPad) vs. **paper field forms** (no drop down menus, prone to data entry mistake)
	Different equipment (rented auto level for one of the crews)
	Broken tape measures (calculation errors associated with “burning a meter”)
Within-crew variation (MV, ESC visit 1 vs. visit 2)	The two field crew members **switched tasks** in each basin between visit 1 and visit 2
	Disturbances from multiple crew visits (kicking substrate, removing substrate after measuring, eroding banks with foot traffic, breaking soft LWD pieces, etc.)
	Weather (attitude, speed of work) and different water heights/flow (zones 1 and 2 fluctuating, different LEW heights, etc.)
Between-year variation (MV, ESC 2014 vs. MV, ESC 2015)	**Natural changes** in sample reaches that occur over time, particularly when there are high flow events.
	Electronic field form (Access file on Panasonic ToughPad) used in 2015 vs. **paper field forms** (no drop down menus, prone to data entry mistake) used in 2014
	The two field crew members may have **switched tasks** between years
	Weather (attitude, speed of work) and different water heights/flow (zones 1 and 2 fluctuating, different LEW heights, etc.)

Precision level of field measurements

- All protocols should contain a specified level of precision for all measurements.
- Precision should be stated on all electronic and paper data forms.
- Precision should not be so high as to not be repeatable.
- The higher cost of precision should be assessed in terms of the time it takes to make a measurement.

Conclusion

Ideally, attribute measurements for status-and-trend monitoring should be consistent (repeatable), precise, accurate, and capable of detecting environmental heterogeneity and change (Roper et al. 2010). The QC analyses in this report provided information on the consistency, precision, between-site heterogeneity, and one-year change. We did not have "gold standard" for accuracy to compare against but relied on our peer-reviewed monitoring protocols and specifically their quality assurance procedures to ensure high accuracy.

The comparison of our QA/QC procedures for field protocols with those reported in the literature for other status-and-trend stream monitoring projects - CHAMP (Bowes et al. 2011), AREMP (Lanigan et al. 2014), PIBO (Henderson et al. 2005) - led us to the conclusion that our procedures are sufficiently rigorous given the project objectives, geographic scale, and budget. Continuing to apply the QA/QC procedures and improving the field protocols as recommended in this document will be sufficient to achieve the desired data quality to characterize status and trends in aquatic and riparian habitat across the OESF.

The comparisons in this report highlight the value of periodically evaluating monitoring protocols having QA/QC procedures, and running annual QC checks. This will not only improve our confidence in the project findings but will allow sharing data with other projects for broader assessments at regional and national scale.

References

Archer, Eric K.; Roper, Brett B.; Henderson, Richard C.; Bouwes, Nick; Mellison, S. Chad; Kershner, Jeffrey L. 2004. Testing common stream sampling methods for broad-scale, long-term monitoring. Gen. Tech. Rep. RMRS-GTR-122. Fort Collins, CO: U.S. Department of Agriculture, Forest Service, Rocky Mountain research Station. 15 p.

Bouwes, N., J. Moberg, N. Weber, B. Bouwes, S. Bennett, C. Beasley, C.E. Jordan, P. Nelle, M. Polino, S. Rentmeester, B. Semmens, C. Volk, M.B. Ward, and J. White. 2011. Scientific protocol for salmonid habitat surveys within the Columbia Habitat Monitoring Program. Prepared by the Integrated Status and Effectiveness Monitoring Program and published by Terraqua, Inc., Wauconda, WA. 118 pages.

Brown, J.D. 2007. Statistics Corner. Questions and answers about language testing statistics: Effect size and eta squared. *Shiken: JALT Testing & Evaluation SIG Newsletter, 12(2*), 38-43 (http://jalt.org/test/bro_28.htm)

Henderson, Richard C., E. Archer, B. Bouwes, M. Coles-Ritchie, J. Kershner. 2005. PACFISH/INFISH Biological Opinion (PIBO): Effectiveness Monitoring Program seven-year status report 1998 through 2004. Gen. Tech. Rep. RMRS-GTR-162. Fort Collins, CO: U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station. 16 p.

Larsen, D.P., P.R. Kaufmann, T.M. Kincaid, N.S. Urquhart. 2004. Detecting persistent change in the habitat of salmon-bearing streams in the Pacific Northwest. Canadian Journal of Fisheries and Aquatic Sciences, 61(2): 283-291.

Lanigan, S., S. Miller, H. Andersen, P. Eldred, R. Beloin, M. Raggon, S. Gordon., S. Wilcox. 2014. Aquatic and Riparian Effectiveness Monitoring Program for the Northwest Forest Plan – 2013 Annual Report. http://www.reo.gov/monitoring/reports/2013%20AREMP%20Tech%20Rpt%20140121%20.pdf

Minkova, T., W. Devine. 2015. Status and Trends Monitoring of Riparian and Aquatic Habitat in the Olympic Experimental State Forest. 2014 Progress Report. Washington State Department of Natural Resources, Forest Resources Division, Olympia, WA.

Minkova, T. and A. Foster. Eds. In prep. Riparian Status and Trends Monitoring for the Olympic Experimental State Forest. Monitoring protocols. DNR Forest Resources Division, Olympia, WA.

Minkova, T., J. Ricklefs, S. Horton, R. Bigley. 2012. Status and trends monitoring of riparian and aquatic habitat in the Olympic Experimental State Forest: study plan.

Minkova, T. and M. Vorwerk. 2014. Status and Trends Monitoring of Riparian and Aquatic Habitat in the Olympic Experimental State Forest. 2013 Establishment Report: Field Installations and Development of Monitoring Protocols. Washington State Department of Natural Resources, Forest Resources Division, Olympia, WA.

Pleus, A.E. and D. Schuett-Hames. 1998. TFW Monitoring Program methods manual for the reference point survey. Prepared for the Washington State Dept. of Natural Resources under the Timber, Fish, and Wildlife Agreement. TFW-AM9- 98-002. DNR #104.

Roper, B.B., J.M. Buffington, S. Bennett, S.H. Lanigan, E. Archer, S.T. Downie, J. Faustini, T.W. Hillman, S. Hubler, C. Jordan, P.R. Kaufman, G. Merritt, C. Moyer. 2010. A Comparison of the Performance and Compatibility of Protocols Used by Seven Monitoring Groups to Measure Stream Habitat in the Pacific Northwest. North American Journal of Fisheries Management 30:565–587.