Watershed Monitoring Project Development Guide for the Voluntary Stewardship Program in Washington

Version 1

(Adopted May 12th, 2022)

Table of Contents

Abstract

Monitoring is an essential component of the Voluntary Stewardship Program (VSP). Effective monitoring is required in order to determine whether counties are meeting the goals and benchmarks defined in each county's VSP work plan. While these goals and benchmarks encompass a variety of themes – such as stakeholder participation, implementation of conservation practices, and critical area functions and values – effective, accurate, and actionable monitoring is a unifying keystone between all themes within VSP.

The approach and content of this guide are intended to provide the "bedrock" of foundational elements relating to monitoring within VSP. The rationale for this approach is that each county's agricultural and ecosystem context is unique, each county defines goals and benchmarks differently, and stakeholders and practitioners at the county-level are presumed to know their watersheds better than those at the state or federal level. The nature, scope, and goals of monitoring endeavors varies across counties throughout the state. Thus, there is not a one-size-fits-all prescription to monitoring that applies to all VSP counties. However, there *can* be a "common language" and "common approach" used for monitoring that is translatable to all VSP counties, despite their differing context and goals.

The intent of this guide is to illustrate these commonalities of approach, language, and scope by leveraging and applying the scientific lens to monitoring – regardless of differences in goals and benchmarks[i] that exist between counties. This guide attempts to communicate these commonalities "from the ground up" by highlighting the essential components required for any type of monitoring. Where possible, additional resources, examples, case studies, and methodologies are provided throughout the guide in order to direct practitioners to approaches most related to their specific goals and benchmarks. This guide is meant to "meet you where you are" in terms of monitoring – recognizing that there is wide variability between VSP watershed work group needs relating to monitoring.

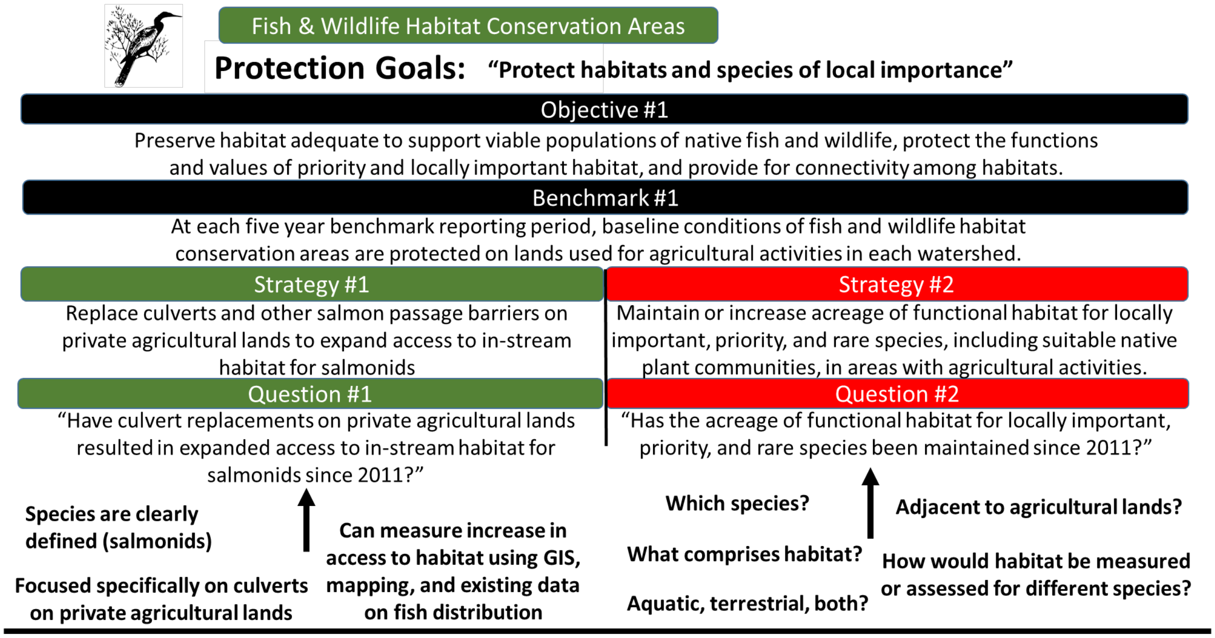
Though the VSP statute encourages the use of existing data, many VSP counties rely on monitoring from other projects conducted by a mixture of county, state, federal, or other organizations (e.g., Conservation Districts, non-governmental organizations, and

[i] Goals and benchmarks in this context throughout this Manual refer to those goals and benchmarks in county VSP work plans. When reference is made to achieving a purpose, the word "purpose" will be used to avoid any confusion with work plan goals and benchmarks.

universities). Typically, these historic or existing projects were not developed with VSP monitoring in mind. Thus, the monitoring information gleaned from these projects may not translate easily to watershed scales, VSP goals and benchmarks or the intersection of critical areas with agriculture activities. Consequently, there are often “mismatches” when VSP groups attempt to operationalize data from these sources for their needs. It is hoped that this guide can help clarify how such problems can be avoided and illustrate how to make the best use of data from a variety of sources for use within VSP. Moreover, the content provided within the guide should help facilitate communication between county VSP watershed work groups and the various partner organizations that conduct complimentary monitoring-related activities throughout VSP watersheds.

A reoccurring theme in this guide can be reduced to four words: “***what is your question?***” All monitoring conducted within VSP should be evaluated in the context of a VSP-related question in need of an answer (Fig. 1). While it is enticing to leverage emerging monitoring tools and data for reporting – especially if they are provided to the watershed work group “for free” – but without a clearly defined question or suite of questions, the scenario becomes “how do I find a question to answer using this free data/tool?” Approaching VSP reporting with data looking for a question will most likely lead down the wrong path.

Figure 1. Illustration of the importance of articulating a clear, meaningful question to orient monitoring activities. Questions that involve generalities, e.g., the example question #2 indicated in red, are likely to result in ineffective monitoring approaches.



Fish & Wildlife Habitat Conservation Areas

Protection Goals: **“Protect habitats and species of local importance”**

Objective #1

Preserve habitat adequate to support viable populations of native fish and wildlife, protect the functions and values of priority and locally important habitat, and provide for connectivity among habitats.

Benchmark #1

At each five year benchmark reporting period, baseline conditions of fish and wildlife habitat conservation areas are protected on lands used for agricultural activities in each watershed.

Strategy #1

Replace culverts and other salmon passage barriers on private agricultural lands to expand access to in-stream habitat for salmonids

Question #1

“Have culvert replacements on private agricultural lands resulted in expanded access to in-stream habitat for salmonids since 2011?”

Species are clearly defined (salmonids)

Focused specifically on culverts on private agricultural lands

Can measure increase in access to habitat using GIS, mapping, and existing data on fish distribution

Strategy #2

Maintain or increase acreage of functional habitat for locally important, priority, and rare species, including suitable native plant communities, in areas with agricultural activities.

Question #2

“Has the acreage of functional habitat for locally important, priority, and rare species been maintained since 2011?”

Which species?

What comprises habitat?

Aquatic, terrestrial, both?

Adjacent to agricultural lands?

How would habitat be measured or assessed for different species?

Consider Figure 1. Within the umbrella of one protection goal and a connected objective and benchmark, there are two strategies with associated questions. The strategy and question for *culvert replacement* (indicated in green) includes sufficient specificity to develop a monitoring strategy. The strategy and question for *functional habitat* (indicated in red) does not include sufficient specificity to develop a monitoring strategy. In this example, the most notable differences between the questions (green vs. red in the diagram) relate to how specific the questions are in relation to *what* will be measured. A purpose of this guide is to help watershed work groups clarify their questions up front so that identifying the types of monitoring approaches and data needed to inform their specific questions becomes easier and outcomes directly show that VSP is having the desired effect of protection and enhancement of critical areas.

Developing a monitoring *plan*, based on the framework established in the guide and references cited within, will help watershed work groups have confidence that the monitoring activities they are conducting, or the monitoring information they are utilizing, will be useful and informative for VSP. Moreover, a clear monitoring plan clarifies expectations of what questions or themes VSP monitoring can conclusively address. A plan ensures that the monitoring questions will be adequately addressed before data is collected. The additional benefits of having a plan are reduced costs, and having a roadmap to help with program continuity as personnel and work group membership changes through time.

The guide is structured in four parts (Fig. 2).

- Volume I focuses on the essential elements of VSP and the key role of monitoring for evaluating VSP goals and benchmarks; it provides a framework for watershed work groups to identify the most important questions to answer with their monitoring program.

- Volume II focuses on operationalizing a monitoring program; it illustrates the different considerations required to develop actionable information to address the monitoring questions identified in Volume I.

- Volume III, which is anticipated to be released in 2022 or 2023, will provide information related to project management, monitoring plan elements, and the types of logistical considerations important for successful monitoring; this section will provide guidance for day-to-day, practical aspects of monitoring to ensure the information collected through monitoring is actionable and aligns to support VSP goals and benchmarks.

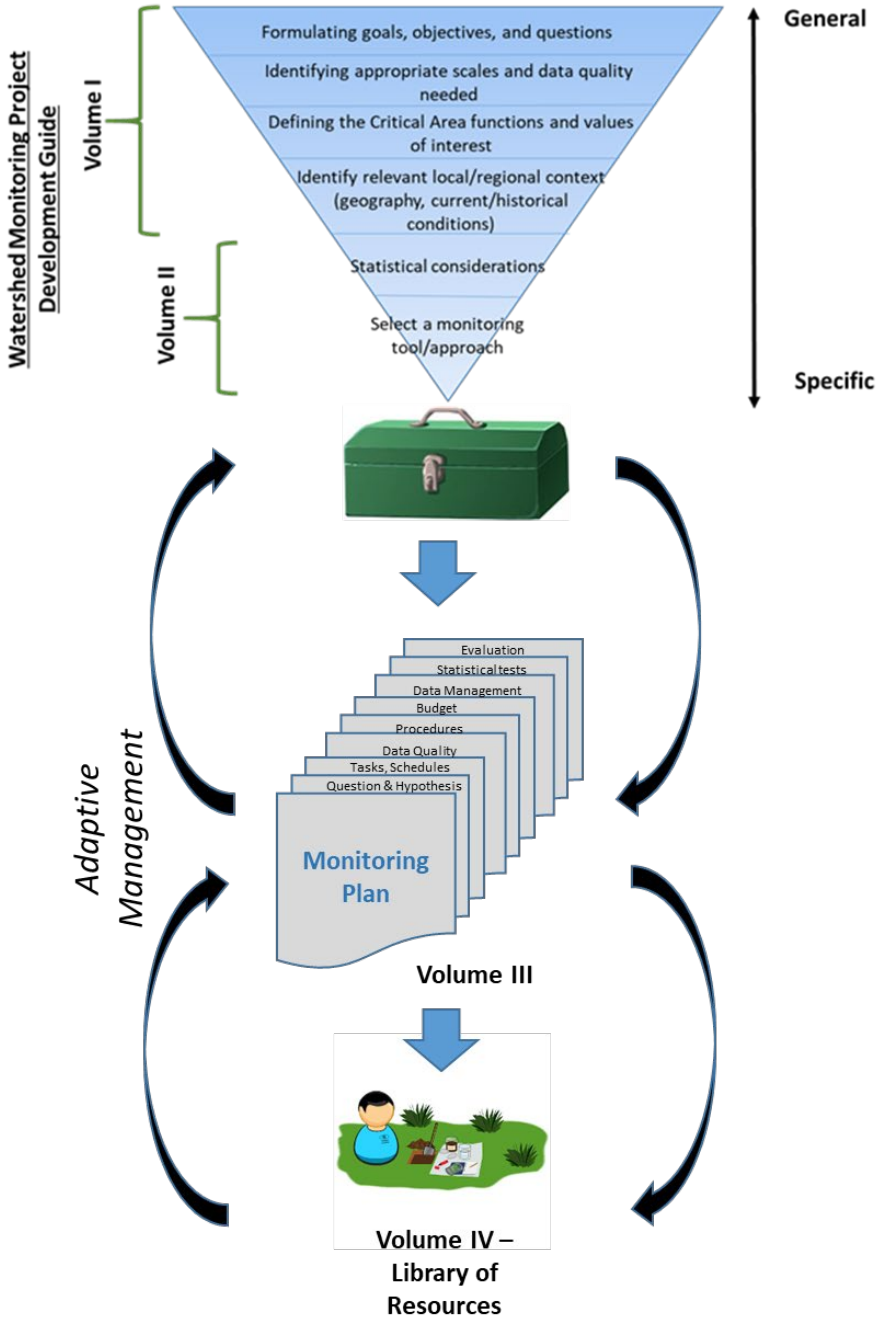
- Volume IV* (Library of Resources) provides VSP-related links and references to documents, data, projects/programs, summaries, as well as guidance and resources for field data collection (e.g., protocols) that pertain to monitoring. Watershed work groups are encouraged to continuously adaptively manage their monitoring programs as questions, conditions, and goals evolve over time. A goal of the **Library of Resources** is to provide stakeholders with information useful for adaptive management.

*Disclaimer: Resources listed in the Library of Resources are for informational purposes only. SCC does not endorse or sponsor the publishers, authors, or organizations that developed the content. The library or resource list intends to provide VSP stakeholders with examples, data, protocols, and documents that support monitoring activities within VSP.

ADAPTIVE MANAGEMENT

Important note: Ongoing adaptive management is needed to align monitoring activities, plans, data, and reporting.

Figure 2. Conceptual diagram of the structure of this guide. The scope moves from general to specific as users are guided through the monitoring development process. Note that the selection of specific monitoring tools, data, and approaches occurs after the necessary foundational themes are addressed. Volume III will focus on developing a monitoring plan based on the foundational themes addressed in Volumes I and II. Volume IV will provide additional guidance and resources for field data collection. Note that ongoing adaptive management is needed to align monitoring activities, plans, data, and reporting.


Watershed Monitoring Project Development Guide
Volume I
Volume II
Formulating goals, objectives, and questions
Identifying appropriate scales and data quality needed
Defining the Critical Area functions and values of interest
Identify relevant local/regional context (geography, current/historical conditions)
Statistical considerations
Select a monitoring tool/approach
General
Specific
Evaluation
Statisticaltests
Data Management
Budget
Procedures
Data Quality
Tasks, Schedules
Question & Hypothesis
Monitoring Plan
Volume III
Adaptive Management
Volume IV – Library of Resources

Background

Context of growth in Washington State and how VSP relates to GMA

Washington is one of the fastest-growing states in the nation. Without careful planning, this increase in people and development could diminish our farmland and other natural resource lands, put people and property at unnecessary risk, and reduce the ability of key ecosystems to function for the benefit of people and other living things in perpetuity .

To that end, Washington State's Growth Management Act (GMA) requires that counties and cities designate and protect natural resource lands (including agricultural lands) and "critical areas" on private and state or locally owned public lands within the entirety of their jurisdictional boundaries. However, VSP applies only on private lands, where agricultural activities and critical areas intersect.

There are five types of critical areas: critical aquifer recharge areas (CARAs), frequently flooded areas, geologically hazardous areas, wetlands, and fish and wildlife habitat conservation areas (FWHCAs). The protection standard under GMA is "no net loss" of critical area functions and values, and this is focused at the tax parcel scale (unlike VSP, which is focused at the watershed scale). GMA gives local jurisdictions regulatory authority for these actions.

VSP is different than regulation under the GMA. VSP is an alternative approach to protecting critical areas where they intersect with agricultural activities.[ii]

- *Voluntary versus regulatory*. VSP is voluntary. That means that a county has a choice to opt-in to VSP, or not.[iii] Participation by landowners in VSP is entirely voluntary – they can choose to participate, or not, at any time. Each county work group must plan for and adaptively manage for landowner participation or the cessation of participation.

- *Watershed scale versus parcel scale*. VSP aims to protect critical areas through voluntary actions on the part of participants, and protection is measured at the watershed rather than parcel scale. VSP work plans are developed by a multi-stakeholder watershed work group, and are based on measurable objectives that are monitored and adaptively managed.

[ii] **RCW 36.70A.710**

[iii] However, once a county opts-in to VSP, the county must abide by certain statutory requirements. Those statutory requirements are beyond the scope of this Monitoring Guide.

- *Agricultural activities versus agricultural land*. The VSP focus is on agricultural activities rather than AG land designations. Agricultural activities include all agricultural uses and practices, as defined in the Shoreline Management Act[iv].

- *Monitoring and adaptive management versus enforcement actions*. VSP requires the county to monitor its work plan goals and benchmarks to ensure those goals and benchmarks are being met. VSP uses adaptive management of county work plans to identify work plan goals and benchmarks that are not being met and provide actions that would meet them.

Counties choose one of two options to meet GMA requirements for protecting critical areas where agricultural activities occur:

1. *Traditional GMA*: Enforce existing parcel-level based critical areas regulations on agricultural landowners. Prior to 2011, this was the only option available to counties.

2. *VSP*: Use the VSP approach to engage landowners with actions they can take voluntarily to protect critical areas at a watershed scale. Twenty-seven of our 39 counties are using VSP.

VSP, through each county's work plan, allows counties and landowners options for protecting critical areas in places where agricultural activity is conducted.

Rather than utilizing regulations and enforcement, counties enrolled in VSP use financial incentives to engage agricultural landowners with voluntary actions that protect critical areas while maintaining the viability of agriculture across the state. A key benefit of this approach is that it allows for locally-led and locally-focused efforts for critical areas protection rather than using a one-size-fits-all, top-down approach, which can be more inefficient and costly.

[iv] **RCW 90.58.065**

Overview of VSP

Program overview

A key difference between the GMA and the VSP approach is that requires for monitoring. Under GMA, it is presumed that Critical Area Ordinances (CAOs) protect critical area *functions and values* by virtue of established regulations and enforcement. However, in VSP, this was done through a three-year planning process undertaken by a local county watershed work group. Each watershed work group crafted a VSP work plan which includes protection and enhancement goals and benchmarks for all five critical area types, as well as goals and benchmarks focused on maintaining agricultural viability. Plans are required to be adaptively managed throughout the county's participation in VSP. All 27 VSP counties have approved work plans and are now implementing those plans. Monitoring is intended to demonstrate ("prove") that goals and benchmarks are being met.

Watershed scale focus

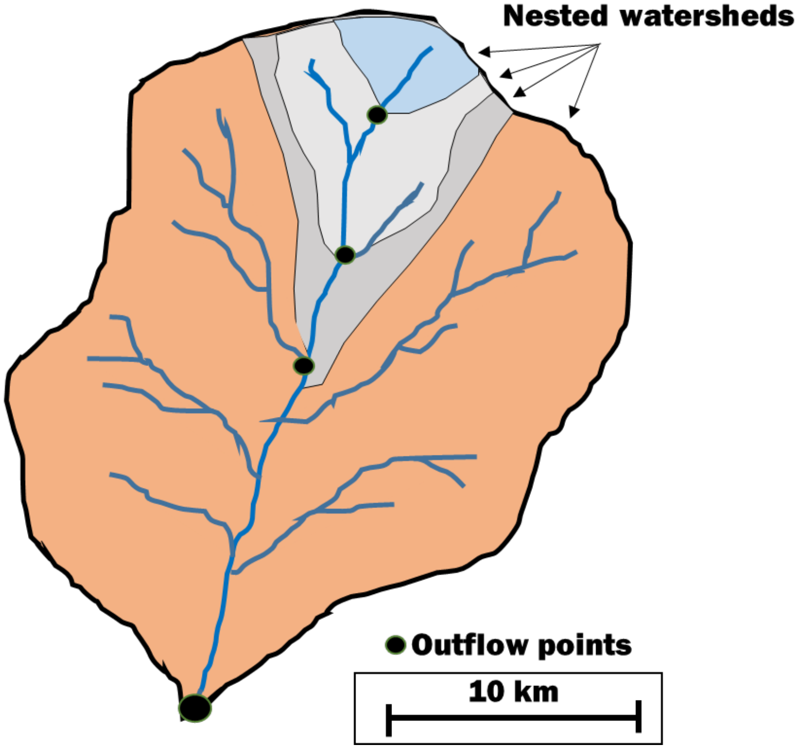
Ultimately, the deciding factor for whether a county can continue participating in VSP is whether critical area functions and values are protected at the watershed scale, as measured against the baseline date of July 22, 2011. Every five years, in a report, each county must assert whether or not it is meeting its work plan's goals and benchmarks. If the county asserts that it is, and the State Conservation Commission Executive Director agrees, then a county may continue to participate in VSP.[v] Information from a suite of monitoring approaches may be integrated into the decisions and activities of each county's watershed work group. Data collected via the various monitoring approaches can inform the county's watershed work group activities and adaptive management efforts to protect and/or enhance critical area functions and values, maintain and promote stakeholder participation, and implement conservation practices.

A watershed is defined as "…a land area that channels rainfall and snowmelt to creeks, streams, and rivers, and eventually outflow points such as reservoirs, bays, and the ocean"[1]. The outflow point selected for watershed delineation will directly affect the size of the watershed. An outflow point at the mouth of a large river will result in a larger-sized watershed, while an outflow point at the mouth of a tributary draining into a large river will result in a smaller sized-watershed (Fig. 3). Thus, a large watershed is composed of a series of nested sub-watersheds[2]. The scale of watershed focus (e.g., sub-watershed vs. larger watershed) depends on the monitoring goals and benchmarks

[v] A complete discussion of the five-year reporting process and discussion is beyond the scope of this Monitoring Guide, but can be found in **Policy Advisory 05-18 – VSP Implementation Reporting – 2 year and 5 Year Reports**

outlined in the watershed work plan and/or questions important to the watershed work group.

Figure 3. Conceptual example of a larger watershed composed nested sub-watersheds. The boundary of a watershed depends on the outflow point chosen for delineation.



County work groups

County work groups were designated when VSP funds were made available in late 2015 and early 2016. The work group is expected to include broad representation of key stakeholders within the watershed. At minimum, this representation should include representatives of agricultural and environmental-oriented groups, as well as tribes that agree to participate. Counties should encourage existing lead entities, watershed planning units, or other integrating organizations to serve as the watershed group.[vi]

Watershed work groups are responsible for developing a work plan to protect critical areas while maintaining the viability of agriculture in the watershed. This group remains responsible for the implementation of the work plan, adaptively managing that plan, and for reporting on the success of the plan to the Commission and county. Practitioners of monitoring within VSP should be focused on carrying out the monitoring activities

[vi] RCW 36.70A.715

outlined in the watershed work plan. If the scope or approach of the monitoring activities outlined in the plan needs to be changed, this should be done in coordination with the watershed work group. Because monitoring information is a statutorily-required and an essential component of VSP reporting, it is crucial that the watershed work group is active in the domain of VSP monitoring within the county.

County watershed work plans

VSP work plans were developed by watershed work groups and evaluated by the VSP Technical Panel, which is composed of members from the Washington State Conservation Commission (SCC), Department of Fish and Wildlife (DFW), Department of Ecology (ECY), and Department of Agriculture (WSDA). Topics addressed in the watershed work plans include: critical areas and agricultural activities, the economic viability of agriculture, an outreach plan for landowner contact, identification of who will provide landowner assistance through VSP, and measurable programmatic and implementation goals and benchmarks.[vii]

The Technical Panel reviewed each county's VSP work plan and made a determination whether the plan, in conjunction with other plans and regulations, would protect critical areas while maintaining or enhancing the viability of agriculture in the watershed. Once the Technical Panel determined a watershed work plan met these criteria, then the SCC director had to approve the plan. All 27 VSP counties have approved work plans. If changes to the watershed work plans are desired, e.g., for monitoring, watershed work groups may make changes via adaptive management.[viii]

A note about Individual Stewardship Plans (ISPs) and confidentiality.

VSP on-the-farm plans (as opposed to VSP work plans, which are county-wide planning documents) are unique to VSP. On-the-farm plans in VSP are called "Individual Stewardship Plans" (ISP's) in the VSP statute, but the statute does not define an ISP.[ix]

As a result, each county decides for itself what an ISP is and how it is defined. Some counties use ISP's simply as a checklist of practices to engage a landowner during outreach. Others use ISP's as full NRCS farm plans.

Each county, through their local county work group and in their VSP county-wide work plan, decides how to use the ISP in their county. There is no requirement in the VSP

[vii] RCW 36.70A.720

[viii] A complete discussion of adaptive management is beyond the scope of this Monitoring Guide. For a complete discussion of adaptive management under VSP, please see Policy Advisory #07-22 Adaptive Management.

[ix] RCW 36.70A.750

statute for ISP's to use NRCS planning standards or process. Therefore, some ISP's will not meet NRCS planning standards because the ISP isn't designed to be similar to an NRCS farm plan. Other counties use ISP's as full NRCS farm plans, and therefore those ISP's will meet NRCS planning standards.

Each VSP county, in their county-wide work plan, has established what an ISP is and how it will be used. A review of a county's county-wide VSP work plan will reveal if the county is using the ISP as a full NRCS farm plan or not, and what the planning requirements will be for that county with respect to ISPs (i.e. do they have to be to NRCS planning standards or not).

The Commission has created a policy advisory[x] to assist those counties in determining if ISP's are confidential. While a full review of that policy advisory is beyond the subject of this Manual, a brief review might be helpful.

Under current Washington state law, all public documents are subject to disclosure under RCW Chapter 42.56 unless exempted in statute.[xi] However, there is an exemption for farm plans developed by conservation districts.[xii] RCW 89.08.560 also deals with farm plan disclosures and includes a definition of "farm plan."[xiii] However, RCW 89.08.560 only applies to farm plans developed by conservation districts – other entities who may be technical service providers under VSP wouldn't have this statute apply to them.

The terms "voluntary stewardship plan" and "individual stewardship plan" are used interchangeably in the VSP statute. Neither term is defined. ISPs that meet the same criteria as the "farm plan," as described in RCW 89.08.560, are confidential and exempt from disclosure, unless permission is granted by the landowner or operator.[xiv] There could be additional confidentiality requirements if a landowner's farm plan is a requirement under federal law.[xv]

[x] Policy Advisory 01-17 – Confidentiality of VSP Individual Stewardship Plans

[xi] RCW 42.56.070(1)

[xii] RCW 42.56.270 (17)(a) (b)

[xiii] RCW 89.08.560(2)

[xiv] RCW 42.56.270(17)(a); RCW 36.70A.720(1)(g) and (i); RCW 36.70A.750; RCW 36.70A.755.

[xv] See Section 1619 of the Food, Conservation, and Energy Act of 2008 (the 2008 Farm Bill), which prohibits disclosure of certain information by the Department of Agriculture (USDA) and its cooperators, now codified in 7 U.S. Code § 8791 - Information gathering.

Goals and benchmarks

Goals and benchmarks are established within each watershed work plan, and are designed to meet the statutory requirements of VSP.[xvi] Goals are the end toward which effort is directed. They identify what the project is trying to accomplish and what the end product will be (e.g. the critical area functions and values to protect). Goals are defined by watershed work groups for programmatic, natural resource, and economic resource themes. Programmatic goals address progress on implementation of the work plan – including landowner participation and stewardship plan implementation. Natural resource goals address the types of critical areas to be protected or enhanced. Economic resource goals relate to the protection or enhancement of agricultural viability within the watershed.

Benchmarks are specific metrics used to establish trends and conditions that relate to the goals outlined in the watershed work plan. Benchmarks serve as a point of reference from which measurements may be made (e.g., number of acres of a specific conservation practice or critical area type). Benchmarks are developed by watershed work groups to assess whether the overall goals established in the watershed work plan are being met.

Each VSP Work Group must conduct a self-assessment every five years, focused on whether or not they are meeting their goals and benchmarks. They are required to submit this information in a report to the state. Each member of the Technical Panel reviews the 5-year reports and decides whether to agree, partly agree, or disagree with each Work Group's assertions. Each agency's assessment, along with the county's five-year report, is provided to the SCC Executive Director for their consideration.[xvii] These reports utilize information from each county's various monitoring approaches to provide evidence demonstrating whether goals and benchmarks are being met. *Thus, clear, robust, and accurate monitoring information is crucial for evaluating each county's VSP goals and benchmarks.*

Protection and Enhancement of Functions and Values

In the context of VSP, the terms "protection" and "enhancement" have specific meanings that relate to monitoring. *Protection* refers to the prevention of the degradation of critical area functions and values existing as of the VSP baseline date, July 22, 2011.[xviii] *Enhancement* refers to improvement of the processes, structure, and

[xvi] See RCW 36.70A.720

[xvii] See RCW 36.70A.720 (2) (b) and (c)

[xviii] RCW 36.70A.703 (8)

functions existing as of July 22, 2011.[xix] In other words, if, at the watershed scale, functions and values have not degraded relative to the baseline of July 22, 2011, then protection goals are being met for that watershed. If they have degraded relative to the baseline, then protection goals have not been met. If functions and values have improved in the watershed relative to the baseline of July 22, 2011, then enhancement is occurring.

It is important to remember that VSP focuses on the watershed scale, not the parcel or land-owner scale. While VSP is based on monitoring at the watershed scale, monitoring data used to support VSP reporting may be collected at a smaller scale – such as a stream reach, field, or sub-watershed – and then aggregated to report at a larger watershed scale. In a given watershed it is possible that activities on some parcels will contribute to the degradation of functions and values while others may improve functions and values (Fig. 4). For VSP, what's of most importance to know for each watershed is whether the "lift" of agricultural practices that improve functions and values (e.g., conservation practices) is enough to outweigh the "drag" of agricultural practices that may degrade functions and values (e.g., conventional practices or polluting activities). Monitoring within VSP is required to demonstrate what direction the "natural resource dial" is moving (Fig. 4) to answer the following question: "is the sum of land treatments in a watershed having a positive, negative, or neutral (no change) effect on functions and values relative to the baseline of July 22, 2021?" If the answer to this question is that the sum of land treatments results in no change to functions and values, then protection goals are being met. If the answer is that there is a positive change in functions and values, then enhancement is occurring. If the answer is that negative change is occurring, then protection goals are not being met.

Note that implementation goals and benchmarks, which specify the type and amount of conservation practices to be implemented in each VSP work plan, are estimates of how the totality of practices within a watershed are expected to affect critical area functions and values, e.g. either protecting or enhancing functions and values. Effectiveness monitoring (see *Monitoring* section below) is used to determine the actual effects of implementation on functions and values. Thus, meeting "protection" or "enhancement" goals or benchmarks for implementation does not, on its own, demonstrate that functions and values were protected or enhanced in the absence of confirmatory effectiveness monitoring. Effectiveness monitoring is required to demonstrate trends in functions and values over time.

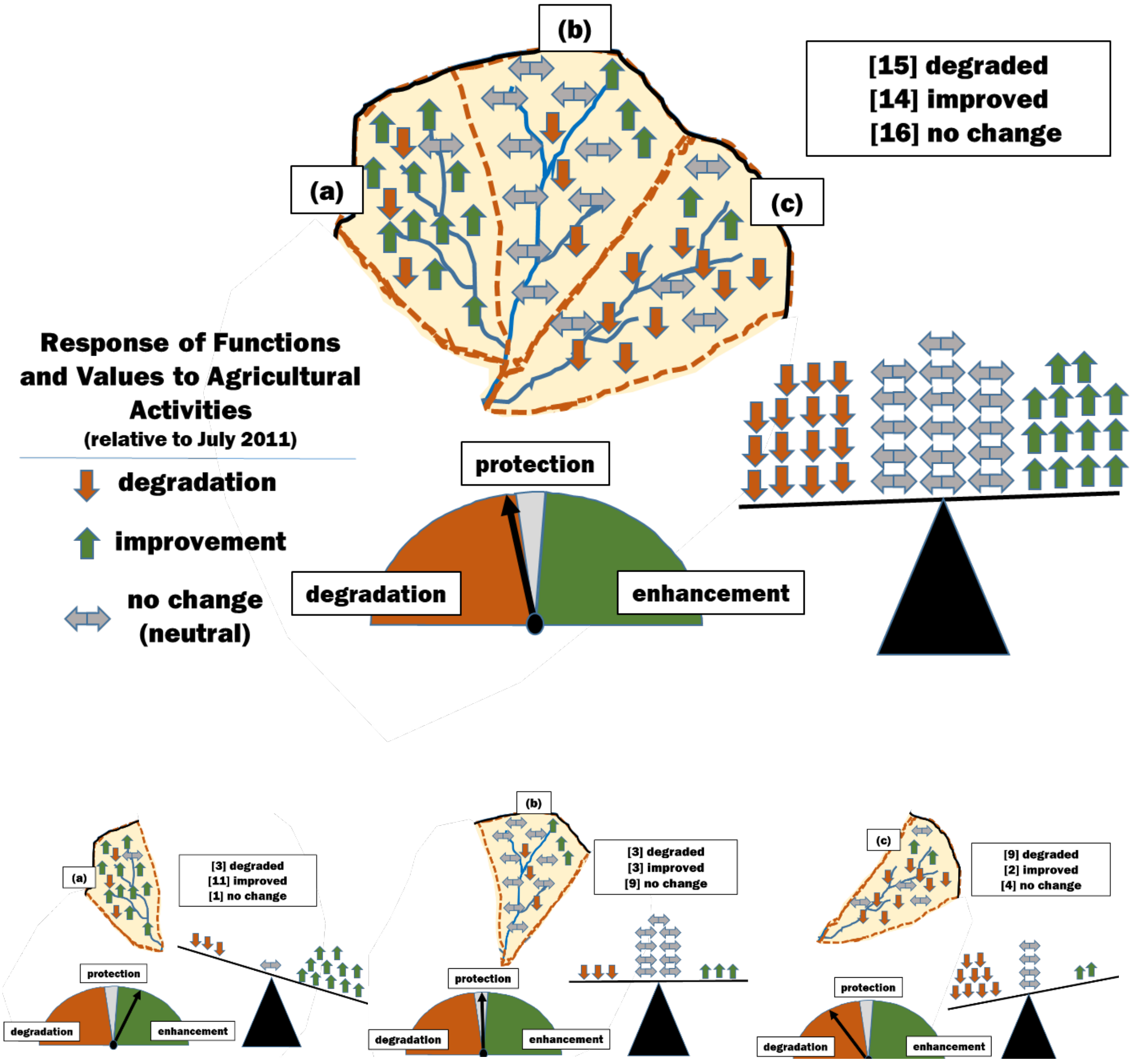
[xix] RCW 36.70A.703 (4)

Refer to the graphic in Fig. 4 – at the scale of the larger watershed, overall functions and values have declined slightly. This pattern is driven primarily by the degradation of functions and values in sub-watershed *c*. Prioritization efforts may focus on implimenting land treatmenst or BMPs within sub-watersheds *b* and *c* – where there are more opportunities to change the trajectory of functions and values – and maintaining land treatments or BMPs within sub-watershed *a,* where improvements in functions and values are already occuring. The graphic in Fig. 4 is used simply as a conceptual example. Actual functions and values measured in a watershed will depend on the goals and benchmarks outlined in the watershed work plan. Ultimately, decisions relating to the prioritization of implimentation and monitoring are up to each county's watershed work group.

The consequences of not meeting VSP goals are different based on whether it is a protection or an enhancement goal. If after 10 years, protection goals and benchmarks have not been met, the county risks failing out of VSP, whereas if the enhancement goals and benchmarks have not been met, "the watershed group *must determine what additional voluntary actions are needed to meet the benchmarks, identify the funding necessary to implement these actions, and implement these actions* ***when funding is provided.***"[xx] Although current funding has not been provided to enhance critical areas, both protection and enhancement goals and benchmarks are included in each county's watershed work plan.

[xx] RCW 36.70A.720 (2) (c) (iv), emphasis added

Figure 4. Conceptual example of a larger watershed composed of three sub-watersheds (*a*, *b*, and *c*), each with 15 “units” of agricultural activities affecting critical area functions and values. Relative to the baseline of July 22. 2021, agricultural activities have degraded, improved, or not changed functions and values – as denoted by the arrow symbols. Note that the overall status of each watershed (enhancement, degradation, or protection) is determined by the combined effects of multiple agricultural activities on functions and values.



Monitoring

Monitoring is a key element of VSP. In contrast to the GMA, monitoring in VSP is statutorily-required.[xxi] It is also required in order to determine whether voluntary actions are, in fact, protecting critical area functions and values. The monitoring program for each county is described in each county's approved VSP work plan. Broadly, VSP monitoring can be used to address at least three VSP-related themes:

- **Stakeholder Participation Monitoring** – assesses the level of engagement of agricultural producers with respect to the goals and benchmarks of the county's VSP work plan. Participation monitoring is required to demonstrate that sufficient numbers of agricultural producers are participating to achieve implementation goals and benchmarks.
    - Example: tracking the proportion of producers participating in VSP within a county and/or watershed using producer surveys, Individual Stewardship Plans (ISPs), or interactive mapping applications.

- **Implementation Monitoring** – tracks implementation of conservation practices (i.e., BMPs) across the landscape within a county and/or watershed, with an emphasis on whether BMPs were installed to proper specifications, when and where BMPs have been implemented, and whether BMPs are being maintained over time. Implementation monitoring is required to demonstrate the amount (i.e., acreage, linear feet) and type (i.e., nutrient management, habitat management, conservation tillage) of conservation practices that are occurring throughout a watershed.
    - Example: field-verification of cover cropping or animal exclusion fencing to ensure practices are to appropriate specifications, accurately reported, and still functioning as anticipated.

- **Effectiveness Monitoring** – determines the effect of agricultural activities on critical area functions and values. Effectiveness monitoring is required to demonstrate whether conservation practices, e.g., at the amount and of the type specified in the implementation goals and benchmarks – are having the intended effects on functions and values.
    - Example: tracking stream turbidity over time in relation to erosion-related conservation practices in a watershed (e.g., no-till or conservation-tillage BMPs vs. conventional tillage); is the sum of BMPs in an area affecting critical areas in the way we expected?

[xxi] RCW 36.70A.720

Monitoring approaches vary by county, but generally include a combination of information from a variety of sources. Sources include, but are not limited to:

- **County organizations** – e.g., a county's health, planning, water resources, or public health departments.
    - Example: collaborating with a county health department to utilize drinking water data to evaluate agricultural-related nutrients in groundwater.

- **State organizations** – e.g., WDFW, ECY, WSDA, or the WA Dept. of Health
    - Example: using information from an ECY TMDL study[3] to estimate the effects of agricultural activities on stream temperature in a watershed.

- **Federal organizations** – e.g., US Geological Survey (USGS), Natural Resource Conservation Service (NRCS), or Environmental Protection Agency (EPA).
    - Example: Utilizing USGS's StreamStats web application[4] to estimate trends in stream flow over time at a specific stream location within a watershed.

- **Tribes** – e.g., the Northwest Fisheries Indian Commission (NWIFC), Upper Columbia United Tribes (UCUT), or individual tribes.
    - Example: utilizing output from an Ecosystem Diagnostic and Treatment (EDT)[5] model to identify priority habitats and limiting factors for salmon and steelhead
    - Example: using "State of our Watersheds" reports to identify priorities for protection and/or enhancement goals[6]

- **Universities –** e.g., Washington State University, University of Washington, or other universities
    - Example: Utilizing the Agland database[7] developed in partnership with the University of Washington to assess changes in the amount of farmland in a county or watershed over time.
    - Example: Participating in a groundwater study conducted by a university.

- **Non-Governmental Organizations** – e.g., American Farmland Trust, The Nature Conservancy, Audubon Society, or other NGO's.
    - Example: using the Coastal Northwest Landscape Conservation Mapper[8] to identify conservation activities that address multiple goals and benchmarks.

- **Conservation Districts –** e.g., one or more Conservation Districts operating within a county.
  - Example: field-based monitoring of conservation practices to determine if they are currently functioning to specifications.

- **Community-Based Monitoring Programs** – e.g., a local river-keeper, high school, or other community-oriented project or program.
  - Example: working with a local high school to collect annual field data on fish counts and habitat quality in priority tributaries.

Problem statement

Monitoring is an essential component of VSP. Because each VSP county has different agricultural, social, and ecosystem contexts, each county defines their goals and benchmarks differently. Thus, evaluating a variety of goals and benchmarks requires a variety of approaches to monitoring. Therefore, ***a one-size-fits-all monitoring approach cannot be successfully implemented for all counties.***

Additionally, because of the integrative nature (e.g., multiple critical area types that may relate to each other; linkages between implementation, effectiveness, and outreach/participation) of VSP monitoring, and the use of data from other organizations to inform VSP activities, there are often "mismatches" when VSP groups attempt to operationalize data from these sources for their needs. These mismatches can confound a watershed work group's ability to utilize monitoring data effectively for managing their activities and reporting on outcomes. County work groups and VSP administrators and implementers must, therefore, analyze the data from other sources to ensure it can be utilized effectively in its monitoring plan for its county work plan. If its utility is limited or ineffective, other data sources should be used to supplement, augment, or replace it.

Another key challenge for monitoring within VSP relates to the balance between obtaining robust, valid, and actionable monitoring data while also respecting the privacy of stakeholders and landowners. The VSP approach relies on evaluation at the watershed scale, not on activities occurring at the parcel (or farm) scale. However, data used to support VSP reporting is still collected at a specific place and time, even if the intent is to aggregate this information to report at a watershed scale. Therefore care is recommended when developing a monitoring plan so that the information gained from monitoring can be used at watershed scales but not unnecessarily impede stakeholder or landowner privacy.

How this guide can help

A central goal of this guide is to help communicate a “common language” and “common approach” to monitoring within VSP. It is anticipated that this common framework will facilitate effective collaboration between watershed work groups and other relevant stakeholders to operationalize monitoring data within VSP. For example, ensuring “everyone is on the same page” is critical to reduce or eliminate mismatches that arise when utilizing data from other programs, organizations, or stakeholders. In developing this guide, the approach was taken to distill elements from a variety of expertise in order to articulate how “the state of the science” best applies to watershed monitoring within VSP.

Recognizing that each VSP county has different levels of familiarity with, and capacity for, monitoring at the watershed scale, it is hoped that by providing the essential elements of VSP monitoring in one guide, watershed work groups can access the relevant information they need to successfully demonstrate that they are meeting their work plan goals and benchmarks. Throughout the guide are examples and references from other resources that may provide more depth on specific topics in order to further support the diverse needs of watershed work groups.

Additionally, care has been taken to address practical aspects of monitoring, such as concerns related to privacy, determining more clearly what types of information is needed to support work plan goals and benchmarks, narrowing the scope of monitoring to align more closely with goals and benchmarks, and referencing sources of information and related organizations that may enhance monitoring with VSP.

Volume I: Deciding why, where, when, and finally how to monitor

Volume I focuses on the key role monitoring plays within VSP. This section provides a framework for watershed work groups to identify the most important questions to answer with their monitoring program. This section also highlights potential pitfalls – and approaches to avoid them – when developing a monitoring strategy, as well as outlining the essential background information required to implement successful monitoring.

Formulating a monitoring question

Generally speaking, “monitoring” refers to watching or tracking something over time for a special purpose. In VSP, one type of “special purpose” is to assess whether critical area functions and values (relative to the baseline date of July 22, 2011) are improving, declining, or remaining the same, at the watershed scale. Similarly, other VSP monitoring types aim to track trends over time – e.g., the level of VSP participation among stakeholders over time, or the amount of conservation practices implemented over time. Each of these monitoring approaches have their own “special purpose” in that the goal is to operationalize information for decision-making by the county work group.

It is important to note that the strength and utility of any monitoring approach is dependent on how well the information gleaned can answer the specific questions a watershed group may have. Specific questions a watershed group may have are driven by the goals and benchmarks in the county-wide VSP work plan. For example, tracking how often a wetland is inundated with water throughout the year may provide answers to questions related to hydrologic function, but may not adequately answer questions related to wetland biodiversity.

Of critical importance is ensuring that a watershed work group’s questions, such as “why are we monitoring?”, are clearly articulated before proceeding with monitoring development. While it is often enticing to jump right into the menu of monitoring protocols – e.g., surface water samples, GIS mapping, fish counts, or rapid assessments – often information collected in such a way does not actually provide a group with desired answers. This is because many of these methods are designed to answer specific questions, and those specific questions may not be the same as the watershed work group’s questions, which are driven by work plan goals and benchmarks for the protection of critical areas and maintenance of agricultural viability.

A preferred approach is to begin with the question(s) and work from there to identify a monitoring approach that can efficiently answer the questions with in a way that is statistically better than random (see *Volume II: Statistical considerations for monitoring within VSP*). Such an approach requires consideration of some key factors:

- "Why are we monitoring?"
  - What is the "special purpose (or purposes)" of the monitoring?
  - Is there a clearly articulated problem statement (i.e., in the watershed work plan or adaptive management plan)?
  - How does your question relate to VSP?
  - Are both the Technical Service Provider (TSP) / monitor and the county work group clear on the specific questions to be answered from our monitoring?

- "Which themes do we intend to inform from our monitoring? "
  - Effectiveness?
  - Implementation?
  - Participation?
  - Some combination of the above?

- "What are we monitoring?"
  - Direct functions and values?
  - Indicators of functions and values?
  - Stakeholder participation?
  - Indicators of stakeholder participation?

- "Where are we monitoring?"
  - Specific watershed?
  - Specific habitat type?
  - Specific land-use?
  - Specific parcels?

- "What level of precision/accuracy do you need?"
  - Do we need a very accurate answer or a just a "ballpark" answer?

- "When will you do something with the monitoring answer?"
  - Under what conditions will the answer trigger changes in projects and activities? (i.e., what is your "do-something-level"?)

- “What scale do we need the answer?”
    - Spatial scale: Sub-watershed? Watershed? County-wide?
    - Timescale: 1-year, 5-year, or 20-year trends? Seasonal? Critical summer period? After a rainfall event?

- “How are we monitoring?”
    - On-the-ground?
    - Satellite imagery?
    - Using models?
    - Using lab samples?
    - Using field observations?
    - Some combination of the above?

- “How do we pay for or access monitoring information?”
    - Local, state and federal monitoring data may be available –
        - Advantages: low or no cost, methodology already provided.
        - Disadvantages: may not answer the questions relevant to VSP, may not be provided in the timeframe you need, may not be provided at the necessary scales for VSP.
    - Use VSP or other funding sources to collect monitoring data
        - Advantages: more control/flexibility over how data are collected to ensure it is informative for VSP, more options for monitoring in ways that are specifically tailored to needs of VSP watershed work groups.
        - Disadvantages: can be time consuming and expensive, may require specialized equipment and staff training.

- “How is the county work group integrating monitoring into the adaptive management decisions that are to be made?”

- “Does the TSP / monitor and the county work group know what actions will be taken in response to monitoring results?”

Carefully reviewing the prompts above and ensuring the watershed work group is clear on the answers is a critical component of formulating a monitoring question. This is especially important given the potential for “mismatches” between available monitoring data and a watershed work group’s stated goals and benchmarks. If a watershed work group can articulate focused, clear answers to the above questions, this will narrow the universe of monitoring approaches to be considered and illuminate pathways for successful monitoring.

Why are we monitoring? Understanding the link between monitoring and your work plan's goals and benchmarks

Effective monitoring requires a clear understanding of the goals the monitoring intends to address. Goals, and benchmarks to reach those goals, are set out in each county's VSP work plan. When problems with monitoring arise, it is often due to a mismatch between the desired goals of a watershed work group and the information made available through monitoring. If watershed groups or stakeholders are not clear on what the project goals are, more likely than not the information gleaned from monitoring will be difficult to interpret.

It is essential to revisit the goals outlined in your county's work plan and consider:

- What critical area functions and values has your watershed work group committed to monitor?
- What does your watershed work group hope to accomplish through monitoring?
- What themes should you focus the monitoring efforts on?
- At what point will the information from monitoring trigger adaptive management actions by the watershed work group? (What's your "do-something-level"?)

In practice, watershed work groups focus on related but different subsets of VSP themes that are of particular importance for their specific context. This is because each VSP county created a VSP work plan to meet the unique needs of protecting critical areas while maintaining agricultural viability in their own particular county. Examples of *variation* between county approaches include:

- Defining critical area functions and values differently, or emphasizing some critical area types over others in relation to specific natural resource concerns.
- Identifying different aspects of stakeholder participation that affect implementation of BMPs in a watershed.

- Categorizing BMPs differently depending upon the functions and values associated with critical areas specific to their county.

For this reason, “successful” monitoring approaches can differ between counties, and may even differ between watersheds in the same county. ***In other words*, *a one-size-fits-all monitoring approach cannot be successfully implemented for all counties.***

While goals and benchmarks may vary between VSP watershed work plans, all VSP counties utilize at least three general types of monitoring themes: effectiveness monitoring, implementation monitoring, and participation monitoring.

Effectiveness Monitoring Goals

Effectiveness monitoring is used by watershed work groups to evaluate agriculture’s impact on critical area functions and values. Each county shares a common need to know whether critical area *functions and values*[xxii] are: (i) increasing, (ii), decreasing, or (iii) remaining unchanged, relative to baseline levels in July 22, 2011. Baseline levels of critical areas functions and values are captured in the county’s county-wide VSP work plan. If a watershed group can demonstrate, using valid data analysis, that functions and values are responding in one of the three ways above, then it should be possible to demonstrate whether VSP is resulting in the protection and/or enhancement of functions and values. Determining whether critical area functions and values are being protected and/or enhanced is a key element of the 5-year reporting process, which reviews and evaluates the success of the county-wide VSP work plan and may cause a county to fail out if goals and benchmarks are not being met. The type of monitoring data needed to evaluate changes to the baseline conditions is dependent on the watershed work group’s goals and benchmarks. For example, if a watershed work group’s goal for Fish and Wildlife Habitat Conservation Areas (FWHCA) is to protect riparian habitat for

[xxii] For a discussion of critical areas functions and values, see *Critical Areas Handbook, A Handbook for Reviewing Critical Areas Regulations*, Washington State Department of Commerce, June 2018, page 12 (“The term “functions and values” refers to the core ecological processes performed by a particular critical area. Critical area functions contribute to the overall health of the ecosystem. Ecological functions of critical areas include flood attenuation, wildlife habitat, water quality, and groundwater recharge,” and page 1 (“Critical areas perform key functions that enhance our environment and protect us from hazards. The beneficial functions and values that critical areas provide include, but are not limited to, water quality protection and enhancement; fish and wildlife habitat; food chain support; flood storage, conveyance, and attenuation (the slow release) of flood waters; groundwater recharge and discharge; drinking water quality and quantity; erosion control; wave attenuation; protection from natural hazards; historical, archaeological, and aesthetic value protection; and recreation. Identifying the functions and values of local critical areas is essential to define the purpose of a critical areas protection program. Each critical area performs different functions and each community assesses the values of the critical areas in its environment differently. Therefore, the purpose of protecting critical areas is unique for each community.”)

salmonid species, the work group will want to include data that tracks changes to the functions and values of riparian habitat on salmonid bearing streams.

Successful monitoring plans have a strong contextual relationship between the information and sources needed to answer the question(s) and the current and antecedent (historic) conditions of a county or watershed(s). It is important to consider that while much of the watershed work group's proposed monitoring will likely occur within a county or watershed boundary, ecosystem processes and functions outside of the local area (e.g., erosion in neighboring counties or upstream watersheds; increased regional wildfire susceptibility associated with global climate change) can influence what is being measured within the county. Moreover, even within a county, variation between watersheds and geography can influence ecosystem functions and values differently. These natural and anthropogenic sources of variability often constrain or otherwise impact monitoring results and methods. It is important to have a clear purpose (i.e., specific questions and/or hypotheses) in order to better clarify which sources and types of information will be needed to meet your work group's objective. There is a balance between including too much information, which can be overwhelming, and oversimplifying information – which can lead to missing key drivers of phenomena in your county watersheds.

In summary, effectiveness monitoring is required to demonstrate whether the amount and type of conservation practices that are occurring on the landscape are having the intended effects on functions and values. Effectiveness monitoring assesses how critical area functions and values are changing over time in response to agricultural activities. In practice, we almost never have a 100% sample of the functions/values we are interested in – water quality, habitat quality, flood storage capacity, etc. Because of this, we use statistics to tell us if our results – e.g., differences we are seeing – are "better than random" estimates of what is happening, because we don't have 100% of the information available (i.e., too much complexity). Content in Volume II of this guide and references cited throughout are intended to provide the foundational approaches necessary to tailor effectiveness monitoring to the specific needs of your county's VSP.

Implementation Monitoring Goals

Implementation monitoring is required to demonstrate (i.e., quantify) the amount, type, and spatial extent of conservation practices that are occurring within the county or its watersheds. Implementation goals should be reflective of how conservation practices are expected to affect critical area functions and values. Implementation of conservation practices is how work plan goals and benchmarks are met. Conservation practice implementation affects, and is affected by, critical area functions and values, and stakeholder participation. Clearly knowing which aspects of conservation practice implementation are in your work plan and why they are important is critical to ensuring continuity in your VSP monitoring activities.

Implementation monitoring has a strong nexus with a watershed work group's goals and benchmarks. If your goals include establishing a relationship between ecosystem functions and values and the number, type, or extent of conservation practices at the watershed scale, it is critical that the way implementation is monitored can adequately inform this overall objective. In practice, this may mean identifying the mechanisms of how a specific conservation practice, or group of practices, may affect the functions and values outlined in the work plan. For example, "upland habitat management" is a conservation practice, but simply reporting that upland habitat management practices were conducted may not shed light on *what* the habitat is, and *which species* the habitat is for. Addressing the aspects of conservation practices that affect specific functions and values (e.g., those in the work plan) will help clarify relationships between patterns of implementation and effectiveness.

There are many approaches to monitoring the implementation of conservation practices. Typically these would involve a combination of mapping in a GIS (e.g., mapping BMP data by watershed using data obtained from the SCC, NRCS, a conservation district, or self-reported by producers) and field checks (e.g., randomly field-checking a percentage of practices to ensure they are installed to specifications and still properly functioning; "windshield" surveys). In contrast to effectiveness monitoring, implementation monitoring typically doesn't require sophisticated statistical tests, such as outlined in Volume II of this guide. It is presumed that the majority of conservation practices being installed are reflected in the implementation data we are receiving and tracking over time. Thus, while some statistical handling of these data may be needed (i.e., "what percent of random field-checks showed that BMPs are still functioning?", "how much error do we estimate in our implementation monitoring data?"), in general, approaches to implementation monitoring will not require *t*-tests and the like.

Examples of implementation monitoring approaches include the Agricultural Conservation Planning Framework[9], projects funded through the NRCS's National Water Quality Initiative[10], or through EPA's summary of conservation tracking tools[11].

Participation Monitoring Goals

Participation among county stakeholders is a critical component of VSP. Participation monitoring is required to demonstrate that sufficient numbers of producers are participating to achieve implementation goals and benchmarks. Inadequate participation may result in not meeting the protection and enhancement goals outlined in each work plan. Participation monitoring is used to track the level of engagement from agricultural producers over time to help evaluate questions about program efficacy, outreach, and marketing. To determine what type of participation monitoring is needed in a VSP county, it may be useful to consider what aspects of participation are most important to ensure the success of VSP in your county. Is the watershed work group concerned primarily with the total number of participating stakeholders? Or, is it important to have proportional representation among different stakeholder groups (i.e., orchardists, cattle ranchers, etc.)? Perhaps it is more important for your group to have participation by stakeholders that implement specific practices or otherwise have a unique effect on the landscape. Ensuring your watershed work group is clear on why participation matters, what type of participation is needed to ensure the goals and benchmarks of the watershed work plan are achieved, and how participation increases or decreases in relation to incentives, will help orient the most effective ways to monitor participation.

Critical area types and functions

Critical areas perform key functions that enhance our environment and/or protect us from hazards. Each critical area performs different functions and each community assesses the values of the critical areas in their environment differently.[xxiii] The beneficial functions and values provided by critical areas include, *but are not limited to:*

Wetlands

- Flood Storage
    - ○ Wetlands may store flood waters during periods of high water runoff, acting as natural tubs that attenuate peak flood and surface water volumes. (WDFW)
    - ○ Surface water that may otherwise cause flooding is stored to a greater degree in wetlands than typically occurs in terrestrial environments. As a result, peak flows in streams and rivers are directly related to the total area of wetlands in the watershed, or to the area of wetlands in the headwaters of the system. Wetlands reduce peak flows in streams and rivers by slowing and storing water in overbank areas and by holding back runoff that would otherwise flow directly downstream and cause more severe flooding. (ECY)
- Water Quality Improvement
    - ○ The performance of the water quality functions by wetlands (i.e., removing sediment, removing nutrients, and removing toxic compounds) depends mostly on the structure of the vegetation that reduces water velocities and causes sediments and pollutants to settle and on the chemical and biological properties of the soil in the wetland. It is the geomorphic characteristics of the wetland and the physical structures found therein that control how a wetland improves water quality. (ECY)
- Shoreline & Erosion Control
    - ○ The ways by which wetlands decrease erosion are somewhat different east and west of the Cascades. This is a result of the differences in the patterns of precipitation and snowmelt between the two areas. For example, in depressional wetlands of eastern Washington, the characteristics within the wetland that decrease erosion are the total storage capacity of the wetland and the relative amount of flow captured from the upgradient contributing basin. The events that cause erosion in eastern Washington are different than in the western part of the state. Summer thunderstorms can cause highly erosive flows at times when most depressional wetlands are dry. As a result, the entire storage capacity of the wetland is usually available to reduce water velocities rather than just the live storage. Depressional wetlands with no outlet store all surface waters coming into them and therefore have the most potential to decrease erosive flows. (ECY)

- Natural Products (food/medicines)
- Food/Habitat for Fish & Wildlife

[xxiii] A complete discussion of the designation of critical areas is beyond the scope of this Monitoring Guide. For further discussion of how local jurisdictions designate critical areas, please see **Department of Commerce's Critical Areas Handbook, Chapter 2 – Resources for Designating and Protecting Critical Areas, June 2018**

    - Scientific literature suggests that size is not a significant factor in contributing to most wetland habitat functions. Rather, habitat structure, connectivity, and wetland hydroperiod are much more significant factors in determining habitat functions than size alone. The literature emphasizes that small wetlands are critically important to amphibians, particularly when connectivity between wetlands and with adjacent uplands is maintained. (ECY)

- Aquifer recharge
    - The characteristics within a wetland that result in the recharge of groundwater are the same for wetlands in both the eastern and western parts of the state. The potential for recharge in a wetland occurs when wetlands hold back precipitation and surface flows to create ponded areas. This ponded water then infiltrates into the groundwater system because of the “head” or pressure created by the depth of water on the surface. If the hydraulic head created by upslope groundwater is greater than the hydraulic head created by the ponded water, recharge will not occur. Groundwater recharge occurs only in a subset of depressional wetlands and some riverine wetlands that impound and hold surface water. Wetland types that do not impound surface water do not have the potential to recharge groundwater. (ECY)

Frequently Flooded Areas

- Flood Storage
- Reduced Erosion/Sedimentation
- Groundwater Recharge
- Hydrologic Connectivity
- Food/Habitat for Fish & Wildlife
- Nutrient/Sediment Distribution
    - Floodwaters transport nutrients that are beneficial riparian communities and aquatic plant life. (WDFW)

Geologically Hazardous Areas

- Erosion Prevention
- Landslide Prevention
- Food/Habitat for Fish & Wildlife
    - Frequently, geologically hazardous areas are too steep to farm or build - yet some of these areas contain important habitat. (WDFW)
- Sediment Input in Streams/Rivers
    - Feeder bluffs are sources of natural erosion and are important contributors of fine sediment to waterways, improving in-stream habitat for many species. (WDFW)

Critical Aquifer Recharge Areas

- Water Quality Improvement

- Drinking Water Provisioning
- Hyporheic Input for Streams/Rivers

Fish and Wildlife Habitat Conservation Areas

- Food/Habitat for Fish & Wildlife
- Migration Corridors
  - Seasonal habitat use, population viability. (WDFW)
- Vegetative Cover
  - Type, composition, and spatial extent. (WDFW)
- Breeding/Nesting Areas & Winter Range
- Riparian Specific Functions
  - Shade, temperature regulation, pollutant removal, wood recruitment, and organic matter input. (WDFW)
- Shrub-steppe Specific Functions
  - Food, cover, migration, winter range, breed/nesting areas. (WDFW)
- Species Specific Functions
  - E.g., raptors reducing rodent populations on vineyards. (WDFW)
- Sustains Native Species/Ecosystems/Biodiversity
- Pollination of Wild/Cultivated Plants
- Predation of Pest Insects by Native Species/Communities
- Societal Food Provisioning (e.g., fisheries)

What are we monitoring? Relationship between functions & values, indicators, condition, quality, and integrity

Overview of terms

Within the realm of “functions and values” and monitoring, many terms and concepts are used, such as “indicator”, “condition”, “quality”, “integrity”, “structure”, and “composition”. Different monitoring protocols (and the organizations/agencies that developed them) may use some or all of the above terms when describing specific monitoring approaches. While the terms above are related in scope – e.g., one may utilize an *indicator* to assess the *condition*, *quality*, or *function* of a wetland – there are nuances and overlap among the various definitions of the terms and these differences in interpretation can affect monitoring. While the basic goal of any VSP monitoring is to determine whether functions and values are being protected or enhanced, or not (i.e., degradation), at the watershed scale, most monitoring approaches do not measure functions and values directly. Monitoring typically evaluates specific elements of functions or values, such as an indicator of a function or the constituent elements of a specific function. Because of this, confusion may arise among practitioners due to the somewhat different uses of various terms between disciplines (e.g., hydrology, engineering, or ecology) or between organizations (e.g., local/state/federal agencies, NGOs, consultants, or universities). The goal of the section is to clarify commonly encountered terminology and help practitioners distill this information for use in their monitoring activities.

Why do these terms matter?

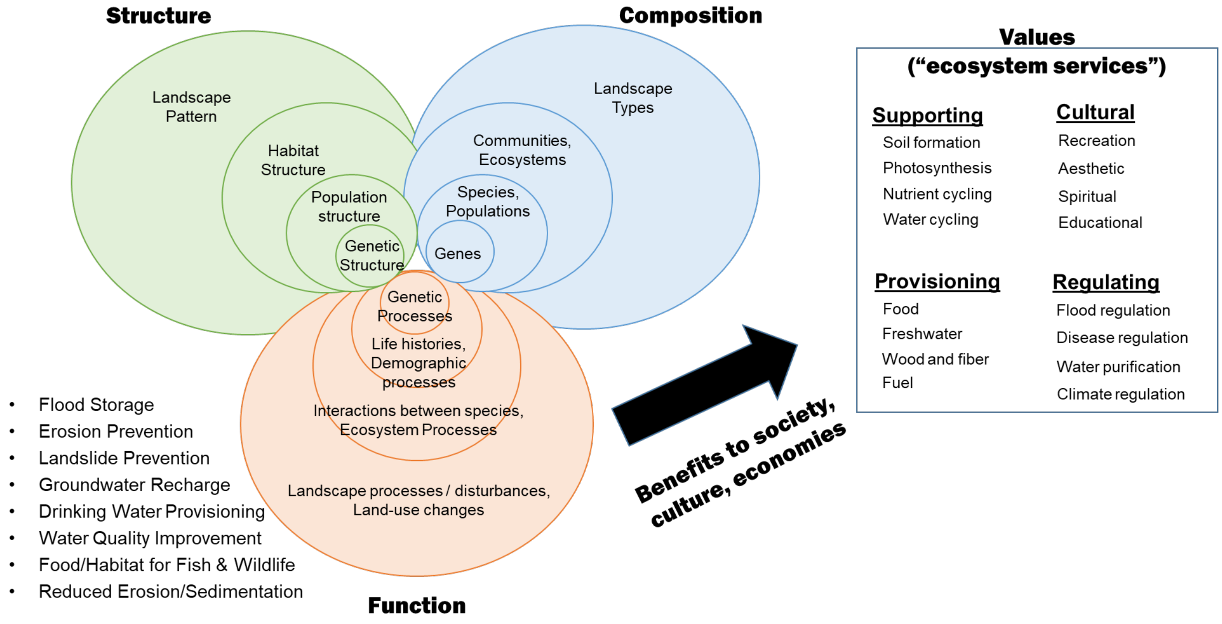
Many monitoring approaches and methods – such as existing "standard operating procedures" from state/federal agencies, NGOs, universities, or others – were developed to address a set of specific questions or objectives. Many of these approaches do not measure *functions* or *values* directly, but they do measure aspects of *structure, composition*, *integrity, quality,* or *condition.* For example, existing field-based assessments conducted by the WDFW, Washington State Department of Natural Resources (DNR), and ECY utilize protocols to evaluate the "integrity" and "condition" of natural resources and critical areas. Examples include WDFW's Ecological Integrity Assessment[12], DNR's Ecological Integrity Assessment[13], and ECY's Wetlands Rating System[14]. While the goal of many assessments is to characterize the *function* of a given critical area, quite often *function* is inferred from an assessment of "integrity" or "condition" – or an *indicator* of function, integrity, or condition. WDFW's EIA defines integrity as "an assessment of the *structure*, *composition*, and *function* of an ecosystem as compared to reference ecosystems operating within the bounds of natural or historic disturbance regimes"[15] (emphasis added). Similarly, ECY's wetlands rating system categorizes wetlands along a gradient (Category I, II, II, and IV) based on wetland functions and special characteristics (e.g., rarity, sensitivity to disturbance, and the ability to replace a given wetland). However, in order to categorize wetlands in this way, wetlands assessments are conducted of various wetland elements relating to *function* (capacity to store water, e.g., from flooding), *structure* (proportion or arrangement of vegetated cover and soils), and *composition* (type of vegetation and soils present) (Fig. 5). The results of an assessment of an individual critical area's composition, structure – and then ultimately function – are often compared to a "reference condition" of the same critical area type to evaluate the relative "integrity", "condition", or "quality" of the critical area (Fig. 6).

Relationship between function, structure, and composition

For purposes of VSP, the focus of monitoring should be to establish whether functions and values (i.e., identified in the county's work plan) are increasing, decreasing, or remaining unchanged. Functions and values themselves are almost always underpinned by some combination of interacting patterns and processes (Fig. 5). For example, flood attenuation by wetlands (i.e., function) is the result of interactions between a combination of specific soils, hydrologic regimes, and/or vegetation (i.e., *composition* and *structure*). In isolation, the specific soil type or the proportion of specific vegetation in a wetland may not be considered a function or value, but the combination of these elements (e.g., soils, hydrology, geomorphology, and vegetation) interacting together may result in a specific function being manifested (e.g., flood storage). In this way, various elements underpinning functions and values are the ingredients that, when combined in a specific way, will result in a specific set of functions and values.

In general, *functions* manifest from interactions between the *structure* and *composition* of living (biotic) and non-living (abiotic) elements of an ecosystem. These interactions often occur at multiple scales from the very small (genes or molecules) to the very large (landscapes) (Fig. 5). In this context, *composition* refers to the identity of the biotic and abiotic elements underpinning a function. Examples include the presence/absence or abundance, density, or concentration of specific soils, geologies, hydrologic regimes, molecules (e.g., dissolved oxygen), genes, species, communities, or ecosystems. For example, wetland habitat for water fowl may require that wetlands are composed of specific vegetation communities (a community of different species of plants), water volumes (specific range of water storage), or food resources (specific insects, fish, or vegetation). In this example, the wetland is “composed” of specific features, much like a recipe is composed of specific ingredients.

Figure 5. Conceptual example of functions manifesting at different scales (genes to landscapes) from a combination of structure and composition. Values are functions that benefit society, culture, and economies.



Structure refers to the way that living and non-living elements are arranged, similar to the relative amount of specific ingredients in a recipe, the order in which ingredients are added, or the way the ingredients are mixed together to make a recipe. Examples include the layering and spatial pattern of soil or geologic deposits, the distribution and arrangement of pools and riffles in a riparian area, or the distribution and arrangement of habitat across a landscape (i.e., landscape connectivity). Different arrangements of these elements will often affect functioning. For example, a fish habitat survey may identify the existence of riffles, pools, runs and glides, as distinct habitat features within a riparian zone that are important for fish (i.e., *composition*). In addition to identifying that these features are present in a location, it is often useful to know something about the relative proportion of these features, or how they are arranged along the river (i.e., *structure*), in order to relate the information to a specific function (e.g., food/habitat for fish & wildlife).

Functions are the resulting processes (e.g., flood attenuation, nutrient and sediment deposition, water quality improvement, etc) that result from unique combinations of the *composition* and *structure* of ecosystem elements, attributes, or features. Typically, assessments of "functions" utilize information about the composition and structure of a critical area to evaluate or infer what level of function the area is providing. For example, rather than measuring the specific amount of water stored in a wetland during a flooding event (i.e., actually measuring flood storage directly), it may be more feasible to measure the size of the wetland, assess its hydrologic connectivity, geomorphic profile, depth, soil and vegetation communities and then infer whether this combination of structure and composition is likely to result in flood storage. For the most part, standardized assessments combine metrics of composition and structure in order to rank or classify a critical area in terms of level of functioning, such as a qualitative ranking of high-medium-low, some type of overall score (e.g., a function score between 1-100), or unique classes (e.g., a Category I vs. a Category IV wetland). Depending upon the goals of your county's VSP monitoring approach, it may be desirable to use existing classifications or to develop unique classifications to better align monitoring outcomes with goals. Practitioners may have more flexibility to develop customized approaches, such as by borrowing elements of existing monitoring protocols, if they are clear on how the pieces fit together to answer the questions of interest from the county work plan. In other words, being able to answer: "what are the combinations of structure and composition that result in the functions we have outlined in our work plan?" will better align monitoring activities with underlying VSP-related questions.

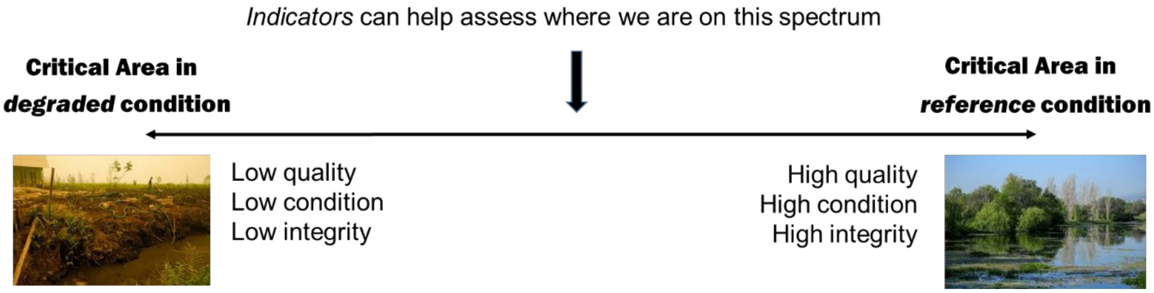
Values and ecosystem services

Values represent the benefits of functions to society, culture, and economies. Often, *values* are considered synonymous with "ecosystem services". *Values* (or ecosystem services) can be categorized by type: provisioning, regulating, cultural, or supporting services/values. Provisioning services include benefits to people that are extracted from nature, such as wood fuel, medicines, drinking water, and food. Regulating services refer to processes that people depend on to regulate natural phenomena, such as flood control, carbon storage, or pollination. Cultural services are those non-material benefits that contribute to the cultural advancement of people, such as recreation, music inspired by nature, or spiritual or philosophical practices that depend on nature. Supporting services refer to the essential processes required for human persistence, such as photosynthesis, nutrient cycling, soils creation, and the hydrologic cycle. More information about values and ecosystem services can be found in the Millennium Ecosystem Assessment synthesis report "Ecosystems and Human Well-Being"[16]

Condition, quality, and integrity

Condition, quality, and *integrity* are largely synonymous, and refer to the status of some focal feature (e.g., wetland, riparian area, aquifer, etc) relative to some reference condition, typically a "natural" condition free from human-induced alterations or disturbances (e.g., critical areas types located in pristine places, or historic conditions of critical areas before human alteration). *Condition, quality*, and *integrity* typically describe how well a focal feature (i.e. critical area) maintains the "natural" structure, composition, and function of similar features in the reference condition (Fig. 6). For example, a diked wetland adjacent to roads would have lower integrity, quality, or condition compared to a wetland located within a pristine nature preserve with no dikes or roads.

Figure 6. Illustration of spectrum of *quality, condition, and/or integrity* in relation to a reference condition. Reference conditions refer to the condition of the critical area in the absence of human alterations, e.g., roads, dikes, or development.



What are indicators?

Indicators are used to measure composition or structure and/or the quality, condition, and/or integrity of some environmental feature, e.g., a Critical Area. There are a wide range of indicators that have been developed for specific purposes. Indicators can be advantageous in that they allow practitioners to narrow the scope of monitoring. Typically, indicators are selected because they broadly represent the condition of an environmental feature. Monitoring specifically for indicators often reduces the complexity of monitoring protocols because the focus on an indicator (if well suited to the question) minimizes the need for monitoring multiple attributes (e.g., composition and structure) of quality, condition, and/or integrity.

Examples of common indicators include *E. coli* (fecal contamination), macro-invertebrates (biological condition of surface waters)[17], or indices of light reflectance from satellites, such as the Normalized Difference Vegetation Index (NDVI; vegetation growth or condition)[18]. Sometimes indicators are utilized in a presence/absence context, such as water tests for presence/absence of bacteria or specific macro-invertebrates that indicate whether a change in condition has likely occurred. Alternatively, indicators can be assessed along a gradient or used to classify conditions into categories (e.g., low-, medium-, or high-integrity).

The use of indicators for monitoring within VSP should be supported by scientifically-supported linkages between the indicator and the functions and values of interest. Thresholds and values of indicators that assess functions and values, but were developed in regions outside of your county, may not translate directly to the monitoring goals of your county's VSP work plan due to differences in local/regional context, ecology, hydrology, etc. Moreover, indicators used commonly in specific regulatory contexts – e.g., the use of fecal coliforms to indicate fecal contamination in shellfish areas – may not actually be effective at identifying drivers of degradation in all contexts. Tests for fecal coliforms often identify non-fecal-related coliform bacteria, which occur naturally in the environment[19]. Thus samples exhibiting high concentrations of "fecal coliforms" may in fact simply reflect samples with high levels of naturally-occurring coliform bacteria that have no relation to fecal contamination. While indicators may be useful for monitoring within VSP, care should be taken when developing or using indicators for monitoring activities.

What level of precision/accuracy do you need?

Determining the **accuracy** and **precision** of VSP-related analysis and monitoring is critical to effective implementation, planning, and reporting.

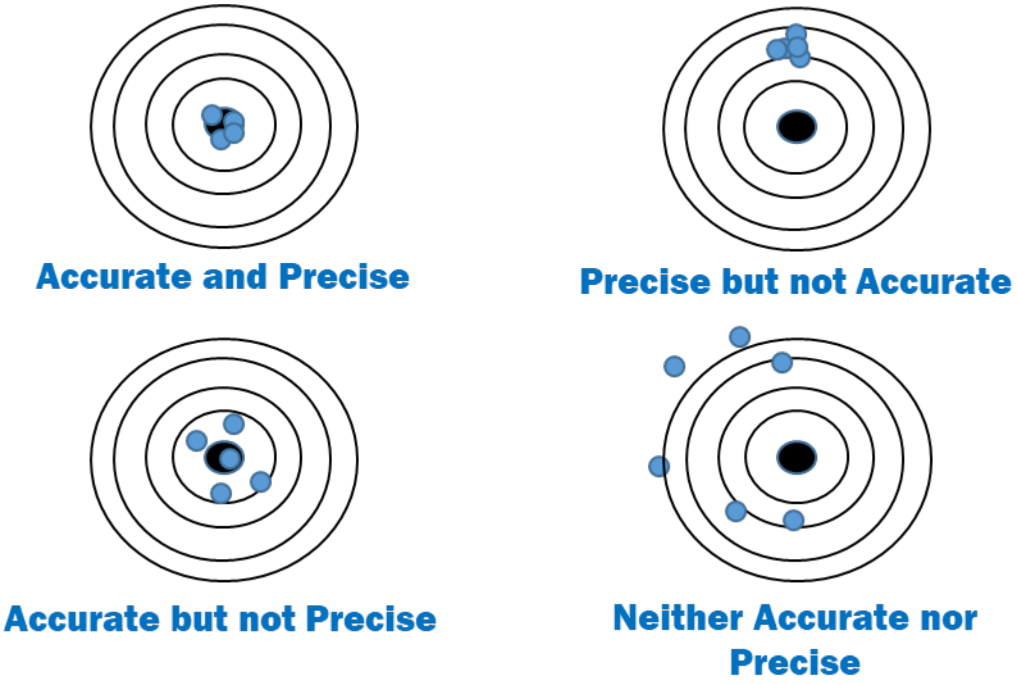
Accuracy refers to how well the analysis, monitoring approach, or observation, reflects the "true" reality that is being measured or assessed.

Precision relates to how consistently measurements of the same phenomenon are similar to each other.

Most people are familiar with the concepts of precision and accuracy in daily life. If you are measuring boiling water (212°F) with a cheap, handheld thermometer, you may note that the temperature reading is not always exactly 212°F. Temperature readings may be within ± 2°F. If you take one or two measurements of boiling water with the cheap thermometer you will likely get a few values between 210°F-214°F. If you take 10 measurements, the average temperature may be very near to 212°F despite the fact that a few measurements were closer to 214°F or 210°F. Even with limited precision, if enough measurements are taken, the "**central tendency**" will approach the true temperature of 212°F. In this case, we could say that the cheap thermometer is accurate, but not very precise.

A more sophisticated thermometer may provide temperature readings within ± 0.1°F. With this level of precision, we may only need one or two measurements for our measured central tendency to approach our “true” value. In other words, low-precision monitoring methods (e.g., a cheap thermometer) may require more samples/observations to estimate the central tendency more accurately.

Moreover, monitoring approaches may also be very precise but still not accurate. Data loggers, pressure transducers, and surface water probes typically require some calibration over time so that the readings reflect “true” conditions. Uncalibrated equipment can often result in very precise measurements that are still not accurate.



Accurate and Precise

Precise but not Accurate

Accurate but not Precise

Neither Accurate nor Precise

Accuracy and precision are key considerations when asking: “What’s your do-something-level’?” In other words, at what level of what’s being measured will there be a change in activities in response? Or, what level will trigger adaptive management actions for your watershed work group? For example, if your watershed work group’s do-something-level is total turbidity of 10 NTU[xxiv] – and current values of turbidity are 7 NTU – then you will likely be interested in a monitoring approach that can at least detect a change of 3 NTU with some level of accuracy and/or precision. In this scenario, using a turbidity monitoring method with an accuracy of ± 10 NTU would likely not provide an adequate answer, because what is desired is to detect a change of as little as 3 NTU.

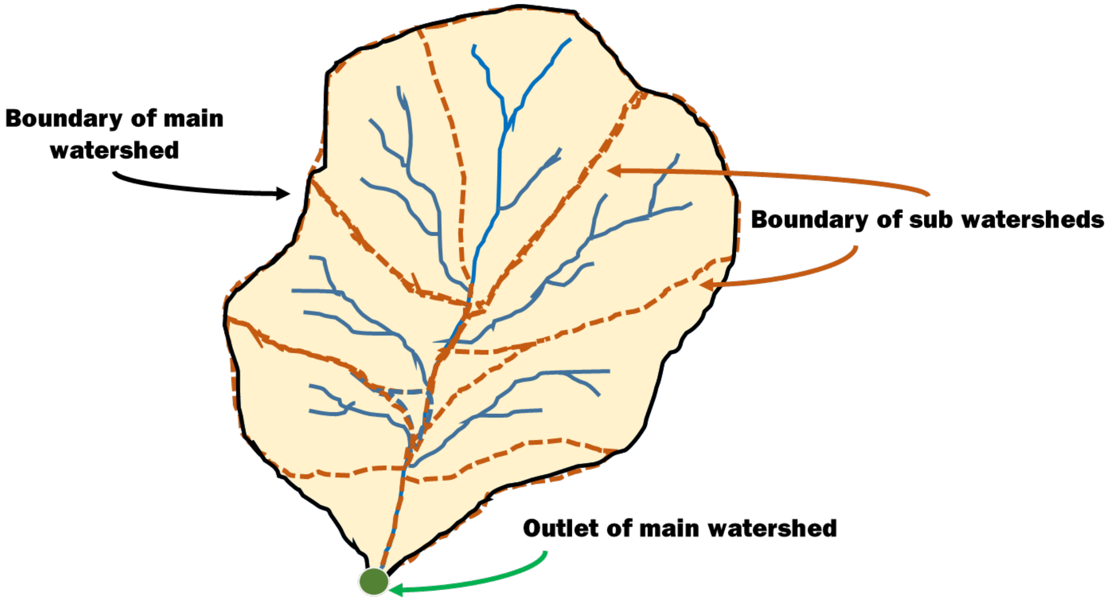
[xxiv] See Definitions of Terms and Acronyms. NTU: Nephelometric Turbidity Units – unit of measure used for quantifying turbidity based on how light scatters in relation to particles in liquid. Larger NTU numbers indicate higher turbidity.

Ideally, monitoring approaches should be tailored in ways that result in actionable information. **Understanding the accuracy/precision trade-offs with respect to the suite of available monitoring approaches will aid your watershed work group's choices of monitoring tools.** More detailed information regarding the influence of accuracy and precision on monitoring outcomes – such as detecting changes over time or confirming a hypothesis – can be found in *Volume II: Statistical considerations for monitoring within VSP*.

Watershed scale monitoring: Considerations related to monitoring and anonymity

One of the primary challenges of monitoring within VSP relates to the balance between obtaining robust, valid, and actionable monitoring data while also respecting the privacy of stakeholders and landowners.[xxv] For the most part, nearly all VSP monitoring approaches leverage information (samples, observations, etc.) from a specific place at a specific time. Even though the focus of VSP monitoring is on the ***watershed scale*** (Fig. 7), watershed-scale data may often be composed of collections of information at smaller scales (e.g., an individual parcel, a stream reach, or an individual ground-water well).

Figure 7. Example of a watershed composed of multiple sub-watersheds. Surface water anywhere in the watershed will ultimately end up at the outlet. The "signal" detected by monitoring at the outlet is dependent upon activities within each of the sub-watersheds.



[xxv] For information on the confidentiality of ISPs, see the Commission's publication Policy Advisory 01-17 – Confidentiality of VSP Individual Stewardship Plans

Since the purpose of most monitoring in VSP is to monitor trends at the watershed scale, in theory the anonymity of individuals within the watershed should be preserved – given that the focus is not on a specific parcel or plot of land, but rather the aggregate of parcels and plots within a watershed. For example, if a watershed is composed of parcels owned by 100 different landowners, watershed-scale trends cannot be directly tied to any specific landowner, but rather to the collective activities of the entire group (Fig. 8a).

In practice, a lot of monitoring information is often collected at a series of locations within a watershed, and then "lumped" together to characterize patterns and processes within the watershed (e.g., total acreage of different types of BMPs; average turbidity across a watershed from 20 different locations; shrub-steppe gains or losses across the watershed estimated from aerial imagery). Different monitoring approaches utilize different focal scales – and some are more appropriate for watershed scales than others. Scaling monitoring approaches and/or data "up" or "down" can be done in order to report at the watershed scale – but if the basic methods are based on aggregating smaller-scale samples to a watershed scale, some thoughtful consideration of how to maintain anonymity of stakeholders is warranted.

Utilizing smaller-scale data (e.g., at the sub-watershed or stream reach scale) may improve the watershed work group's ability to adaptively manage for the protection of functions and values by providing a tighter linkage between specific agricultural activities and specific functions and values. Consider that even within a WRIA-scale watershed, some sub-watersheds within a WRIA may be more erosive or sensitive to nutrient inputs than other sub-watersheds within the same WRIA. Moreover, some sub-watersheds may lack significant agricultural activities (Fig. 8a) – such sub-watersheds may be a lower priority for monitoring within VSP. Monitoring at the sub-watershed scale may provide more actionable information related to which BMPs are most effective or appropriate for a given geography (sub-basin/stream reach) – which in turn can be used internally by the watershed work group for prioritization, outreach, and planning purposes. Sub-watersheds can be aggregated to complete a picture of the larger watershed (e.g., WRIA) and results may be reported (e.g., via the 5-year reporting process) at the larger watershed scale – while smaller scale data at the sub-watershed scale may be used internally by the watershed work group if desired, for purposes of planning, monitoring, and prioritization. One option to consider is "anonymizing" sub-watersheds (e.g., sub-watershed A, sub-watershed B) in order to utilize data at multiple scales while preserving landowner and stakeholder privacy (Fig. 8b).

Figure 8a. A watershed and its sub-watersheds with different levels of agricultural activities. Agricultural activities are occurring primarily in two parts of the watershed, noted in sub-watersheds *c* and *e*. Monitoring surface water at the outlet of the entire watershed will likely detect signals originating in parts of the watershed not relevant for VSP (i.e., no agricultural activities).

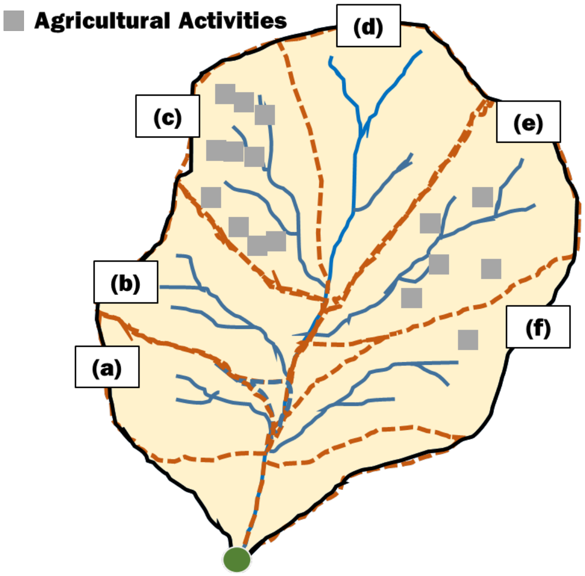
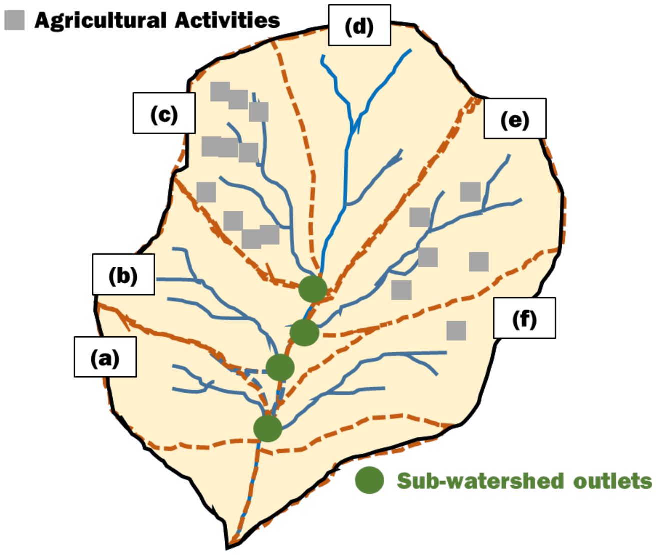


Figure 8b. A watershed and its sub-watersheds with different levels of agricultural activities. Green circles denote the outlets of sub-watersheds. In this example, monitoring surface water at sub-watershed outlets would help isolate the signal of interest (i.e., agricultural activities). Outlets with little or no agricultural activities (locations *a, b, d, and f*) may sampled if there is a desire to compare functions and values between active agricultural sub-watersheds and those with little or no agricultural activities.



Additionally, surface water samples may be collected in such a way as to isolate the potential effects of agricultural activities while reducing the influence of non-agricultural related effects (e.g., urbanization or industrial contaminants). Selecting upstream-downstream pairs of sample units, where the upstream flows through areas known to have agricultural activities, is more likely to effectively monitor the “signals” of ecosystem functions and values most important to your VSP (Fig. 9a). Additionally, it is often desirable to compare watersheds that encompass different conditions – such as different levels of conservation practices or agricultural activities – to assess the influence of these differing factors on functions and values (Fig. 9b). Depending on the size of the sub-watersheds and the scale of the monitoring approaches used, it is possible to monitor so the signal of interest is not related to a specific parcel or plot, but rather a collection of parcels and plots.

Figure 9a. A watershed and its sub-watersheds. Black circles denote sample locations. Upstream (*us*) locations may be paired with downstream (*ds*) locations in order to assess contributions of agricultural activities to ecosystem functions and values.

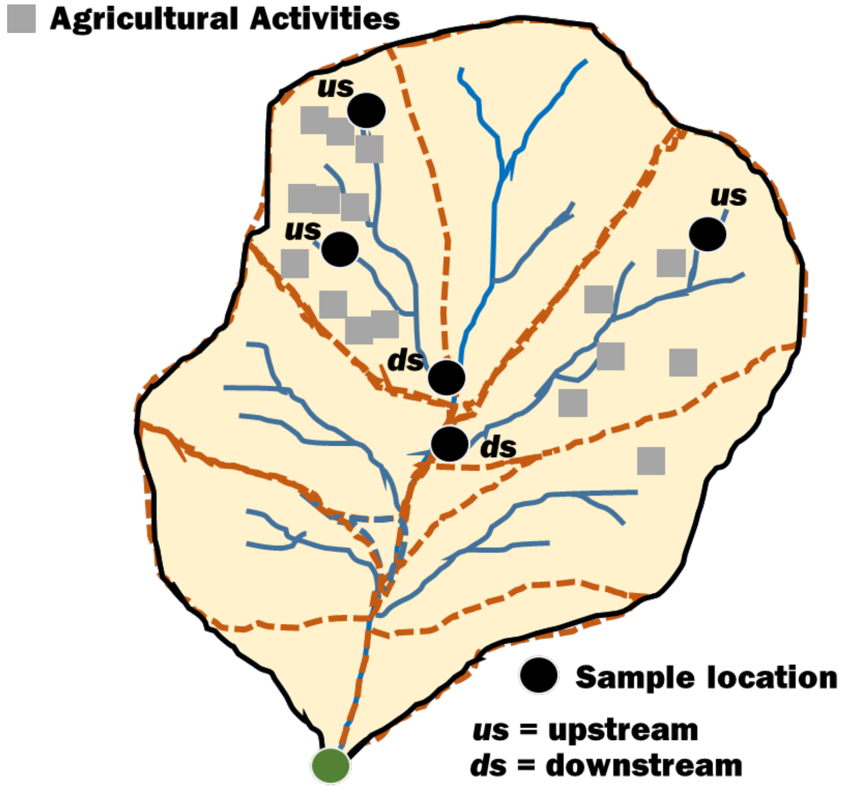
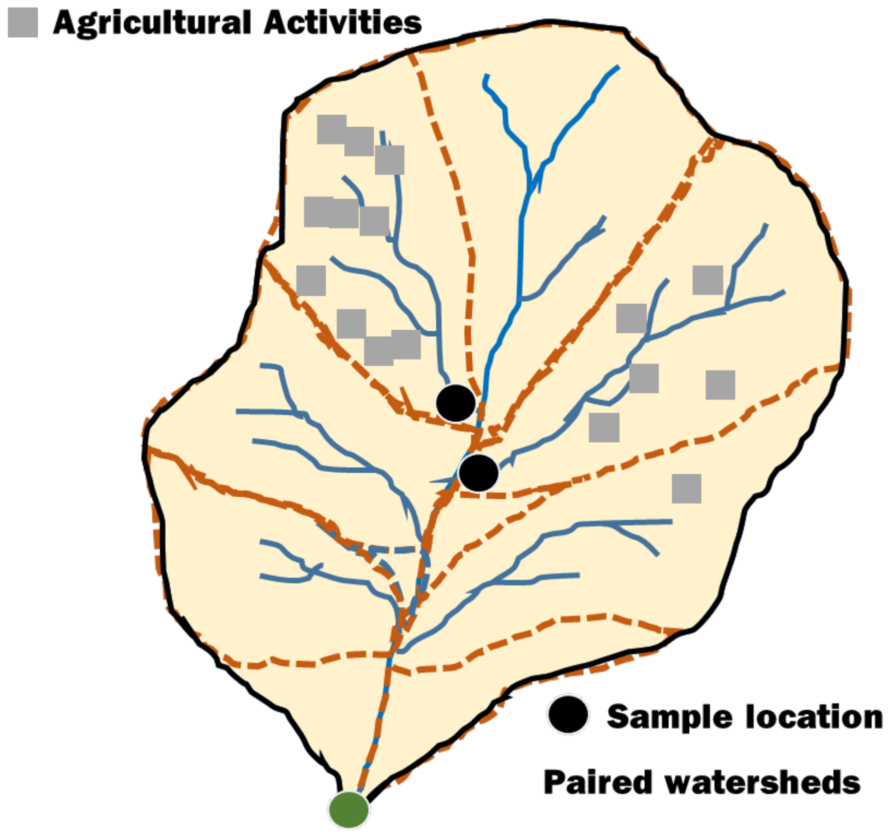


Figure 9b. A watershed and its sub-watersheds. Black circles denote sample locations. Watersheds may be paired to compare how different factors (e.g., conservation practices vs. conventional practices; high no. of agricultural activities vs. low no. of agricultural activities) affect ecosystem functions and values.



Most study designs employ the use of “randomized” selection of locations for sampling, as this a pre-requisite for many statistical approaches (i.e., locations have an equal chance of being selected as a monitoring location, see *Volume II: Monitoring Design - Statistical considerations for monitoring within VSP*). In other words, for the most part, the foundation of monitoring design within VSP is built on random locations (within a region of interest) that together should reflect the general trends and conditions of an area – not to isolate the influence of specific parcels or small-scale locations.

Additionally, the importance of selecting the test, sample, or analysis most relevant to the questions of interest to the watershed work group cannot be overstated. Carefully selecting the monitoring approach and indicators of interest can help ameliorate privacy concerns by focusing the scope of inquiry.

For example, a focus on “sediment” or “cow manure” as a potential driver of the degradation of functions and values may orient the watershed work group to a suite of monitoring approaches. The differing approaches assess a variety of “lenses” with which to view sediment or cow manure, and some of these lenses may affect privacy and anonymity (Tbl. 1). Moreover, the different approaches to monitoring sediment or cow manure have varying abilities to answer specific questions. Note that often “free” data, such as data collected by other organizations within your county, may be oriented toward species or pollutants that are used in a regulatory context – such as fecal coliforms or total suspended solids. For some watershed work groups it may be desirable to use these regulatory-aligned monitoring approaches, such as to ensure data are comparable over time. For other groups it may be undesirable to use regulatory-aligned monitoring approaches because they may be expensive, inappropriate to answer the questions outlined in the work plan, or because of concerns among stakeholders. It is always advisable to ask, and re-ask, “what is the critical area function and concern, and how do we best ask/answer the question?”

If done thoughtfully, monitoring can shed light on the effects of agricultural activities on functions and values while maintaining the privacy and anonymity of stakeholders and landowners within a county’s watersheds.

Table 1. Example monitoring approaches applied two themes, *(a)* sediment in surface water, and *(b)* cow manure in surface water. Note that different approaches inform slightly different questions; some approaches are more common than others in regulatory frameworks.

(a) Theme of interest: sediment in surface water

Monitoring Type	**Monitoring Approach**	**Regulatory?**	**Questions the Approach Can Answer**	**Comment**
Field Observation	Secchi disk	No	How clear is the water?	Informs questions related to clarity of water (of everything)
Field Sample	Imhoff cone	No	How much of the water sample settles as solids?	Informs questions related to erosion, lake or river filling, or gravel bed inundation
Lab Analysis	Total Suspended Solids (TSS)	Yes	How much of the sample is filterable and most likely sediment?	Drying the sample bakes non-soil elements, may not differentiate soil from non-soil solids
Field Sample	Nephelometric Turbidity Units (NTU)	Yes	How much light scatters, 90 degrees to incident beam, due to particles in the water?	Includes sediment and other particles that reflect light

(b) Theme of interest: cow manure in surface water

Monitoring Type	Monitoring Approach	Regulatory?	Questions the Approach Can Answer	Comment
Lab Analysis	Fecal coliforms	Yes	How much bacteria (usually fecal) is present?	Informs questions related to (mostly) fecal-related bacteria - but cannot differentiate between sources (e.g., human vs. cow vs. deer)
Lab Analysis	*E. coli*	Yes	How much of a species known to cause ENT illness is present?	Informs questions related bacteria known to cause *E. coli* enteritis in humans - but cannot differentiate between sources (e.g., human vs. cow vs. deer)
Lab Analysis	Total coliforms	Sometimes	How much bacteria from all sources is present?	Informs questions related to all bacteria present - but cannot differentiate between sources of bacteria (e.g., human vs. cow vs. deer)
Lab Analysis	eDNA or Microbial Source Tracking (MST)	Sometimes	Are specific species of bacteria present?	Requires DNA database/library to reference samples - may differentiate between sources of bacteria (e.g., human vs. cow vs. deer)

How are you monitoring? Different types of monitoring approaches

There are many approaches to monitoring. Each type of approach has associated trade-offs that are important to understand in the context of the watershed group's goals and benchmarks. Most monitoring falls broadly into one or more of the following categories:

Field Observation – observations of conditions on-the-ground, typically by one or more field personnel. Examples include recording the presence/absence of a specific plant or animal, reporting whether a wetland is wet or dry, or categorizing land-use (e.g., grassland or forested).

Field Measurement – measurements of conditions on-the-ground, typically using some type of equipment or probe, such as a staff gauge, or surface water probe. Examples include the use of equipment to measure stream flow, water temperature, or turbidity in the field.

Field Sample – samples of a subset of a population of interest, typically using a quadrat, transect, Imhoff cone, or other standardized approach. Examples include using a series of quadrats or line-point intercepts to estimate plant cover, composition, and/or density in wetlands or rangelands.

Lab Analysis – samples are collected from the field and returned to a lab for analysis. Examples include surface water quality samples (e.g., fecal coliforms or microbial source tracking [MST]) or soil samples analyzed for nutrient content and grain size.

Model – monitoring is based on the results of models, e.g., wildlife population viability or hydrologic models. Examples include Ecosystem Diagnostic and Treatment (EDT) for modeling fish abundance, land-use maps derived from satellite imagery, or projections of groundwater availability based on changing precipitation and temperature patterns.

Each of the above categories of monitoring has associated strengths and weaknesses and should be evaluated by the Technical Service Provider / monitor in conjunction with the county watershed work group to match methods with work plan goals and benchmarks. For example, field observations are typically less accurate and less precise than field measurements or samples, but they can be time/resource efficient if tailored to address a specific question. Lab analyses can often provide high precision, accurate information – but the cost of lab sampling is typically greater than field observations or measurements. Alternatively, models often make use of a variety of other information sources (e.g., habitat conditions, climate, light reflectance, land-cover, etc), sometimes at different spatial and temporal scales, in order to estimate current conditions or future projections.

In practice, combinations of each of the approaches outlined above are often used when monitoring. Ideally, the choice of monitoring approach is dictated by specific questions articulated by each watershed work group in each county-wide VSP work plan. In consultation and communication with your watershed work group, thoughtful consideration of your watershed work group’s questions, how much certainty is desired in the answer to the questions, as well as practical constraints, should guide your choice of monitoring approach.

How local and regional conditions affect monitoring programs

Once a work group has thoroughly articulated the specific purpose of monitoring by distilling down to a key set of questions, a crucial next step is to collate and synthesize the relevant context within which the county and associated watersheds are embedded (Fig. 2). *Much of this information has already been put together in each county's approved VSP work plan*. Integrating key factors across themes, such as land-use, climate, hydrology, and historical conditions, allows the county work group to "connect the dots" between agricultural activities and critical area functions and values in a way that is meaningful to, and specifically tailored for, stakeholders within each county. Ideally, this information is not simply listed in a table, but rather is truly operationalized and integrated into adaptive management and monitoring protocols.

Several contextual themes of particular relevance to VSP are:

Geographic area and surroundings of the county and/or county watersheds – underlying geology, geomorphology, precipitation, temperature, hydrology, and ecology.

Example: counties with watersheds that contain more erodible soils on steep slopes will likely have higher levels of background turbidity in stream systems – potentially influencing sampling frequency and/or "do something levels" for stream turbidity. If your work group's do-something-level is total turbidity of 30 NTU – and current values of turbidity are 25 NTU in watershed A, but 10 NTU in watershed B – then you will likely be interested in more accurate/precise measurements in watershed A compared with watershed B, because you are hoping to detect a smaller difference in watershed A. As a result, your group might decide to sample more frequently or use methods that are more precise in watershed A.

Agricultural and ecosystem context – focal agricultural activities and surrounding ecosystem type(s).

Example: dryland vs. irrigated agricultural activities may focus on different aspects (e.g., timing) of soil moisture for maintaining agricultural viability – this may result in altering soil sampling protocol, intensity, and/or scope; suites of BMPs used for dryland farming may differ from irrigated agriculture and may affect critical area functions and values differently. The type of ecosystem the county is embedded in drives the types of critical area functions and values of importance to stakeholders – e.g., higher abundance of wetlands west of the cascades than east; potentially different functional role of wetlands east of the cascades (i.e., water availability in the summer) than west. If your county work plan's goal is to assess hydrologic connectivity of wetlands, you might consider a different suite of tools than if goals pertain to wetland biodiversity.

Primary natural resource concerns – current or historical natural resource concerns highlighted from previous projects and studies; unknown phenomena.

Example: Previous projects have employed Limiting Factor Analysis (LFA)[20] to elucidate key drivers of fisheries-related habitat loss and degradation in each WRIA across the state. Specific locations and causes of natural resource concerns have been identified for remediation or restoration. Despite the existence of previous natural resource inventories and projects (e.g., LFA), likely gaps in knowledge remain (i.e., "known unknowns"). Understanding these gaps can help direct future monitoring needs or inform how likely an existing monitoring approach will answer the work group's questions given uncertainty. For example, your work group might use monitoring data collected from previous projects in your county to estimate how frequently you should sample surface water (see *Volume II: Monitoring Design - Determining how many samples you need to answer your question: Power analysis*). In watersheds without such data, your group may consider a small pilot project to help inform future sampling intensity.

"Connecting the dots" – linking monitoring themes to evaluate success

Monitoring is a key element of VSP and is statutorily-required. Monitoring is required in order to determine whether the voluntary implementation of conservation practices is protecting critical area functions and values. Monitoring is most useful when the various themes – effectiveness, implementation, and participation – are integrated and the results operationalized for management. Consider that:

(i) Effectiveness monitoring is required to demonstrate that the amount, type, and extent of conservation practices occurring on the landscape are protecting or enhancing the functions and values of critical areas at the watershed scale. Effectiveness monitoring must assess functions and values (e.g., fish and wildlife habitat, wetland flood storage capacity, etc) over time to assess changes relative to the baseline of July, 2011.

(ii) Implementation monitoring is required to demonstrate the amount, type, and extent of conservation practices that are occurring on a landscape (e.g., no. of acres of specific BMP's in a given watershed). Implementation monitoring (and the prioritization and planning of BMP's) should be directly related to the functions and values that are to be protected/enhanced, as outlined by the county work plan.

a. Example: nutrient management practices are implemented in areas/watersheds where nutrients are a major driver of the degradation of critical area functions and values. The amount of nutrient practices implemented is commensurate with the scale of the nutrient problem.

b. Example: tillage practices (no-till or reduced till) are implemented in areas/watersheds where erosion is a major driver of the degradation of critical area functions and values. The amount of tillage practices implemented is commensurate with the scale of the erosion problem.

(iii) Participation monitoring is required to demonstrate that sufficient numbers of producers are participating to achieve the implementation goals and benchmarks outlined in the work plan. In many cases, it may be important to identify key constituencies of producers for specific goals or objectives – e.g., understanding levels of participation of ranchers, orchardists, or grain producers depending on specific natural resource concerns (dissolved oxygen in surface waters vs. reduction in wetlands habitat), key drivers of ag-related effects on critical areas (nutrient management vs. tillage practices), and location (high priority watershed vs. low priority watershed for a given natural resource concern).

(iv) Information gleaned from effectiveness monitoring, implementation monitoring, and participation monitoring should be evaluated and used by watershed work groups to gauge whether or not VSP-related activities are having their desired effects – i.e., are functions and values being protected and/or enhanced? Is sufficient implementation occurring to meet goals and benchmarks? Is sufficient participation occurring in order to meet the levels of implementation needed to achieve goals and benchmarks?

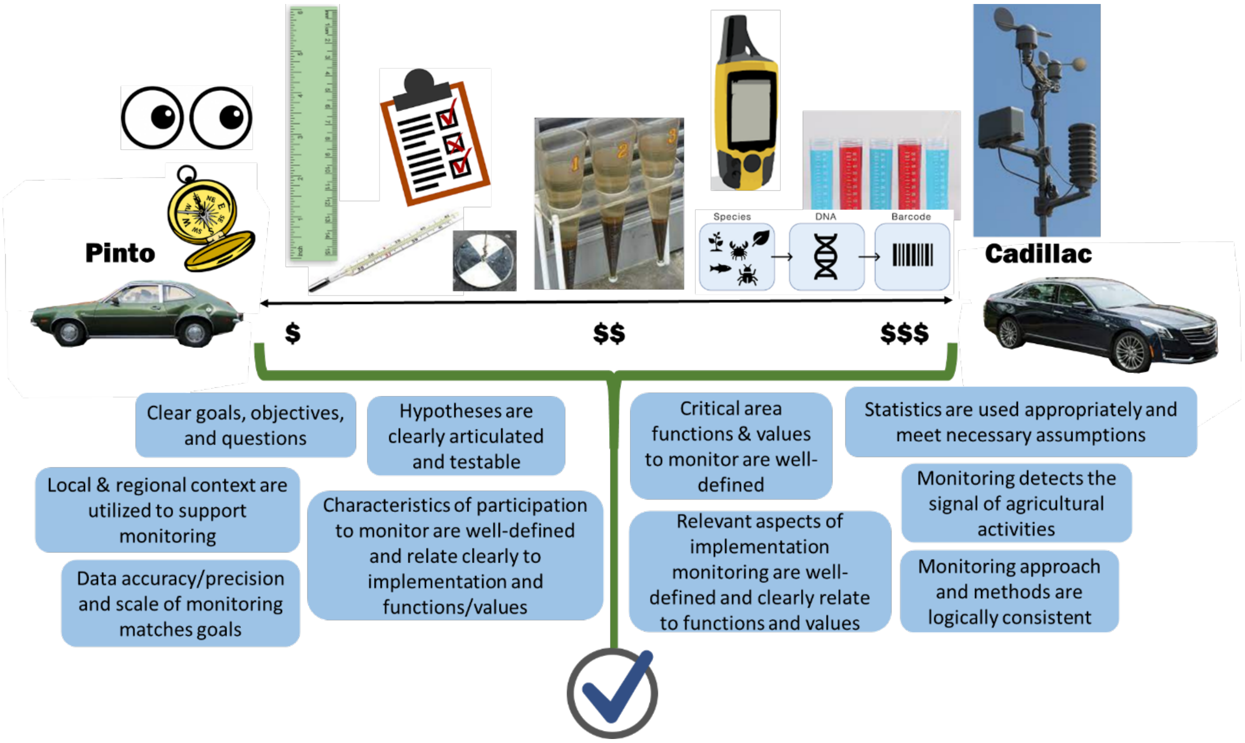
(v) Work groups are encouraged to evaluate monitoring data in "real time" in order to effectively adaptively manage their county's VSP. The power of good monitoring data and information is manifested when data is incorporated into management decisions.

a. Example: implementation monitoring reveals that a large proportion of nutrient management practices are being adopted in watershed A – which has minimal nutrient-related natural resource concerns – while watershed B has severe nutrient-related natural resource concerns but has low levels of nutrient management implementation. Based on this information, the work group has increased outreach to producers in watershed B to increase nutrient management implementation in order to address the nutrient-specific natural resource concerns occurring there.

b. Example: effectiveness monitoring reveals that fish and wildlife habitat is being lost due to agricultural activities in watershed A. Based on this information, the work group has increased outreach to producers in watershed A to increase implementation of BMP's that would restore habitat (e.g., upland habitat management) and/or protect habitat (e.g., livestock fencing). Participation monitoring will demonstrate if outreach is successful. Implementation monitoring will demonstrate how many, where, and what type of habitat-related BMP's were implemented. Effectiveness monitoring will demonstrate if these activities are resulting in the protection or enhancement of habitat within the watershed.

A note on monitoring programs: from “Pinto” to “Cadillac”

Across the 27 VSP counties there are a variety of levels of capacity for monitoring. Some counties have a number of large-scale projects from related agencies that include “Cadillac”-like monitoring approaches utilizing specialized equipment and staff. Other counties have limited opportunities to include these types of monitoring data in their VSP reporting because their counties may not have the luxury of these larger-scale projects and programs. Consider that most agree a Cadillac is a more expensive, higher-end vehicle compared to a Pinto. However, both a Pinto and a Cadillac have the same ability to transport you from one place to another if they are maintained and properly fueled. So it is with monitoring – it is not necessary to use expensive equipment or methods to achieve the objective of demonstrating that work plan goals and benchmarks are being met. What is essential is that the core ingredients of monitoring are operationalized in light of your watershed work plan’s goals and benchmarks, and that these ingredients are incorporated in a logically consistent way. Watershed work groups are encouraged to consider a range of monitoring options without worry that a Pinto-esque monitoring program is not sufficient. If the foundational principals of monitoring outlined in this guide are understood and followed, the monitoring program should be sufficient.


Pinto
Cadillac
1
2
3
Species
DNA
Barcode
$
$$
$$$
Clear goals, objectives, and questions
Hypotheses are clearly articulated and testable
Critical area functions & values to monitor are well-defined
Statistics are used appropriately and meet necessary assumptions
Local & regional context are utilized to support monitoring
Characteristics of participation to monitor are well-defined and relate clearly to implementation and functions/values
Relevant aspects of implementation monitoring are well-defined and clearly relate to functions and values
Monitoring detects the signal of agricultural activities
Data accuracy/precision and scale of monitoring matches goals
Monitoring approach and methods are logically consistent

Volume II: elements of a monitoring plan

Volume II focuses on operationalizing a monitoring program by applying science-based principals to improve efficiencies and outcomes. This section illustrates a range of considerations – from developing hypotheses to estimating the number of samples needed for conclusive results – required to develop actionable information to address the monitoring questions identified in Volume I.

Monitoring Design

Monitoring goals

It is of critical importance when designing your monitoring plan that your work group has a clear understanding of your work plan's goals and benchmarks and how monitoring is intended to help meet those goals and benchmarks. For more information about how to determine which questions are important for your watershed work group to address, review *Volume I: Formulating a monitoring question*. Questions to consider with your work group include:

- What are primary reasons for conducting monitoring?
- Is there a clearly articulated question/problem statement (i.e., in the watershed work plan or adaptive management plan)?
- Are both the TSP / monitor and the county work group clear on the specific questions to be answered from our monitoring?
- Which themes do we intend to inform from our monitoring?
  - Effectiveness?
  - Implementation?
  - Participation?
  - Some combination of the above?
- How is the county work group integrating monitoring into the adaptive management decisions that are to be made?
- Does the TSP / monitor and the county work group know what actions will be taken in response to monitoring results?

Revisiting questions such as those listed above and incorporating them into your work flow will strengthen your work group's ability to operationalize monitoring results to make prudent decisions.

At a basic level, most questions within VSP will fall within one of three categories:

Are/Is _______ improving over time?
Are/Is _______ declining over time?
Are/Is _______ staying the same over time?

The blank spaces above could be replaced with themes related to: critical area functions and values, stakeholder participation, or BMP implementation.

Perhaps the most robust and effective manner to answer these questions, regardless of their formulation, is through the use of statistics. Using a statistical framework allows stakeholders, reviewers, and third parties outside of VSP the ability to "see" what direction the system is going in a way that is "better than random". Statistics can help clarify uncertainty and frame results with levels of confidence. Moreover, effective use of statistics can also help improve efficiencies in projects and programs[21].

Hypotheses

A useful approach to develop statistical support for your monitoring needs is to consider using hypotheses. Hypotheses are informed explanations that can account for observations we make in the natural world. Most often they relate to the effect of some cause, mechanism, or process (e.g., increased precipitation) on patterns we note from observations (e.g., higher soil moisture).

For example, we can hypothesize that an increase in precipitation will result in higher soil moisture based on our experiences observing that soil appears wetter after continued rainfall. In a scientific method framework, we can test this hypothesis by collecting data in some thoughtful way – perhaps measuring soil moisture with a probe before, during, and after rainfall events – and then test whether soil is in fact wetter in response to rainfall.

An elegant way to do this is to test against the *null hypothesis*. The null hypothesis essentially states that the observations or patterns observed are *not* related to the cause you have articulated, but in fact are caused by randomness, measurement error, or some other unknown cause. In our example with rainfall, our null hypothesis could be: increased precipitation *does not* result in higher soil moisture. Our alternative hypothesis, then, would be: increased precipitation *does* result in higher soil moisture.

Once we have defined our hypotheses, the null and the alternative, we can employ appropriate statistical approaches with our data to see how much confidence we have to reject the null hypothesis, and accept the alternative. Typically 95% confidence is the standard for most analysis. For example, after employing our statistical tests we may find that there's a 95% chance that soil moisture increases with rainfall, and only a 5% chance that the data we collected are no better than random, affected by sampling error, or caused by something else besides rainfall.

In light of the example above, more general hypotheses for a work group may look something like:

- Overarching question –
  - "Are critical area functions and values increasing over time?"

- Null Hypothesis (i.e., baseline we are testing against) –
  - "There *has not* been an increase in critical area functions and values" since 2011.

- Alternative Hypothesis–
  - "There *has* been an increase in critical area functions and values" since 2011. In other words, the alternative hypothesis is what you think has happened, but you are using statistics to show you whether you can reject the null hypothesis – and thus accept the alternative hypothesis.

Basically, statistical tests are set up assuming the differences we are seeing (e.g., turbidity between time periods) are based simply due to random chance, and that there is in fact no actual difference in reality.[22,23] Then, using data collected from monitoring, we can test to see how much confidence we have in rejecting the null hypothesis. This process underscores the importance of understanding the relationship between central tendency, noise, and P-values (probabilities). These concepts will be revisited frequently throughout this document.

Statistical considerations for monitoring within VSP

Understanding your data: Trends, noise, and errors

For most statistics within VSP, developing a firm understanding of the concepts of **central tendency** and **noise**, and how they apply to your monitoring program, will be tremendously helpful with respect to creating actionable monitoring information. Generally speaking, more (thoughtful) sampling will result in higher confidence in central tendencies and less influence of noise on results.

- **Central tendency/average (mean)***
  - What are we seeing most commonly? What is the average temperature of the stream? On average, how many salmon returned in our stream system this year? What is the average number of acres utilizing conservation tillage in our county?

- **Noise/variability (standard deviation/variance)***
  - How messy, noisy, or variable is the phenomenon we are measuring? How much fluctuation is there in our measurements of stream temperature? Are our surveys filled out consistently and thoroughly, or do some surveys report detailed information while other surveys have missing information?

- **Probability (P-value)**
  - How likely are we to get a result like this based on random chance? Extremely unlikely? 50-50 chance? 1-in-1000 chance? Lower P-values indicate that what was observed is less likely to happen by chance. Typical P-values to confirm a hypothesis are $P < 0.1$ or $P < 0.05$ (<10% or <5% chance our results were random).

*Note: remember that more *accurate* monitoring methods will better reflect the true central tendency/average, and more *precise* monitoring methods will reduce noise associated with the equipment, method, observer, and model. Increasing accuracy and/or reducing noise typically results in more confidence in your results (i.e., lower P-values).

One of the ways in which practitioners gauge "is this information better than random?", in relation to monitoring data, is through the use of statistical tests and P-values. Statistical tests provide a platform to estimate some outcome (are things the same or different?), and/or level (what is the average number of fish?), and typically provide a measure of confidence in those reported levels. P-values are a common approach to assessing how likely the results were simply "by chance". That is, a low P-value (usually $P < 0.05$), suggests that it's highly unlikely the results were simply random chance. A high P-value ($P > 0.5$) indicates the opposite – that there's a high likelihood the data is not better than random.

In practical terms, most consider a P-value of $P < 0.05$ to be sufficiently robust threshold to conclude that the results of a statistical test are attributable to a specific cause (e.g., turbidity decreased over a 5-year period) rather than random chance. Generally, values of P < 0.1, but P > 0.05, are considered *weakly* significant (Tbl. 2).

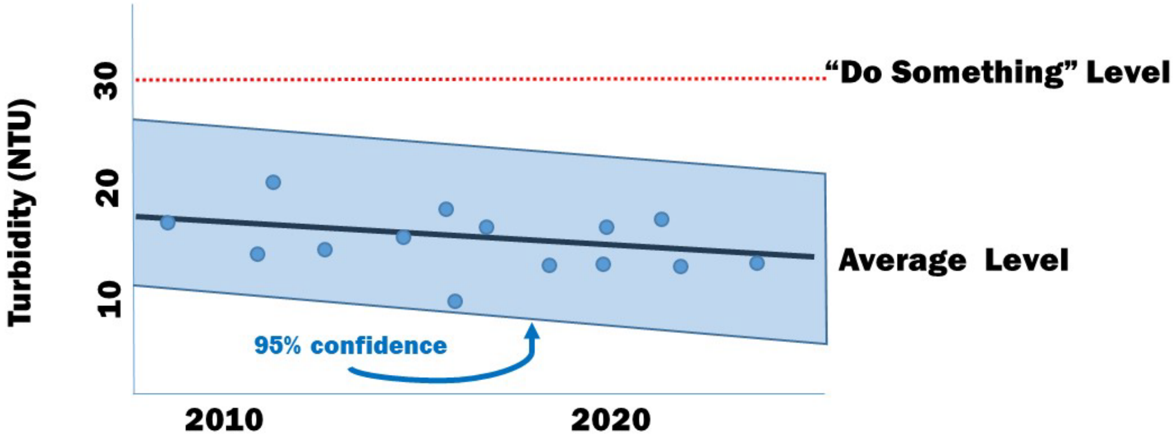
Table 2. Different thresholds for rejecting or accepting hypotheses.

P-value	Interpretation	Inference
P > 0.1	1-in-10 chance or greater the differences were random (>10%)	No significant difference
0.05 < P < 0.1	Between 5-in-100 and 10-in-100 chance the differences were random (5-10%)	Weakly significant difference
P < 0.05	Less than 5-in-100 chance the differences were random (< 5%)	Statistically significant difference

Example of Central Tendency, Noise, and Randomness for Surface Water Turbidity

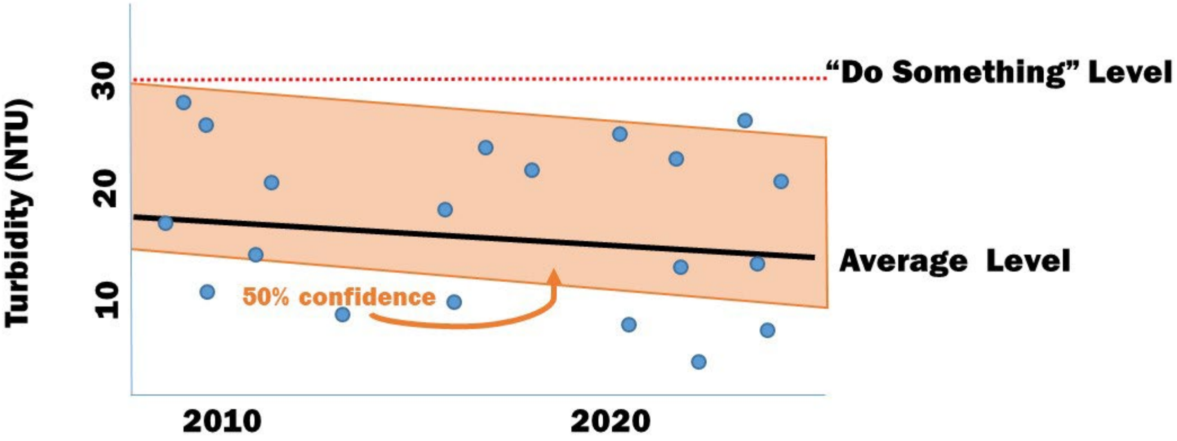
Let's use an example. If a work group's "do something level" for surface water turbidity is 30 NTU, and the results of monitoring show that the central tendency (e.g., average) value is between 5-25 NTU with 95% confidence, then it becomes clear that the true value of turbidity is most likely below the do-something-level of 30 NTU (Fig. 10). This is because there is only a 5% chance that the true value is actually below 5 NTU or above 25 NTU (and thus a 95% chance [confidence] the true values is between 5-25 NTU).

Figure 10. Exploring data trends relative to a "do something" level with 95% confidence.



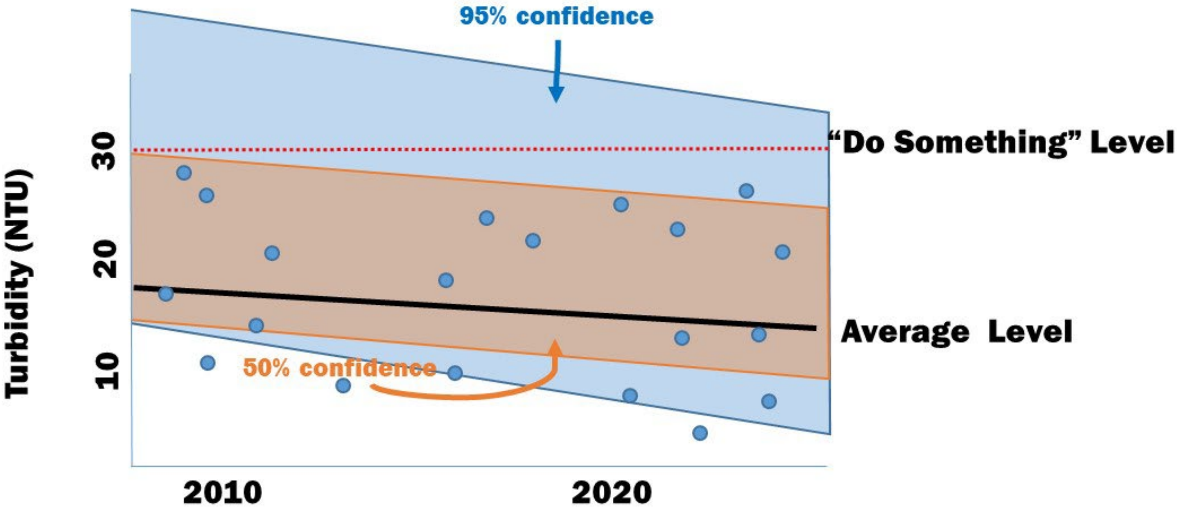
In the same scenario, if the data is noisy (less precise), we may determine that the true value of turbidity is between 5-25 NTU with 50% confidence – but 50% confidence is a 50/50 coin flip (Fig. 11). In other words, at 50% confidence, there's a 1-in-2 chance the actual level of turbidity is below the do-something-level. This is not a high level of confidence. If the data were able to be recollected, it's as if you flip a coin to know whether turbidity is at the do-something-level or not.

Figure 11. Exploring data trends relative to a "do something" level with 50% confidence.



Using the same data set, we find that we can calculate a 95% confidence interval, but this interval is between 2 NTU – 40 NTU (Fig. 12). Our do-something-level is 30 NTU, which is within the 95% confidence level. In other words, we are 95% confident the true turbidity level is as high as 40 NTU, but it could also be as low as 2 NTU – so, we have 95% confidence that our actual turbidity level is either above or below "do something level". Thus we don't have sufficient information from our monitoring to decide to do-something or not-do-something.

Figure 12. Comparing data trends relative to a "do something" level with 50% and 95% confidence.



Discovering that our data isn't sufficient to answer our question – in this case, "should we do something?" – is quite common for the types of data used in VSP, e.g., water quality, habitat assessments, etc. A helpful approach to monitoring is to use existing data, even if insufficiently robust to reduce noise or provide conclusive answers, to estimate how many samples would be needed to be able to more definitively determine whether the central tendency has reached the do-something-level. This data can be operationalized to estimate how much more sampling would need to be done to provide a conclusive answer – allowing work groups to more efficiently target sampling. If no previous data on your topic of interest exists, pilot studies – which are typically designed to be short and exploratory – may be used to collect a small initial set of monitoring information that can be used to estimate future sampling frequency (see *Power Analysis*).

Data Exploration

A critical piece of operationalizing your monitoring data to inform project activities is simply *exploring* your data. Before conducting statistical tests, or looking for trends in our data, we should spend a little time reviewing the underlying data.

Why should we explore our data? Well, in a basic sense, we want to know that the data we are using for decision-making is good quality information. If someone gave you a tank of gas to use for your car, the cautious among us would want to know a few things about the gas inside the container before we put it in our car. Is it regular gas or diesel?

Does it smell like gas? Is it clear? Does it have debris in it? If we just put it in our car without exploring it, as above, it might work fine, or it might damage our vehicle.

The same basic cautions should be applied to our monitoring data. Even if we receive the data from a trusted source, there may be nuances to the information that are essential to know. Some examples include:

- Outliers in the data: extreme values (high or low) can change our interpretation of "true reality" by biasing our results. If one out of 20 readings of our thermometer is 10 times hotter than the other readings, e.g., because we made a mistake writing it down, or the thermometer is failing, our "signal" (average) will appear much warmer than the true temperature. These outliers are fairly easy to see if you plot your data (e.g., Box and Whiskers plots)

- Units of the data: sometimes data sets contain measurements or values in different units (e.g., feet vs. meters or Most Probable Number [MPN] vs. Colony Forming Units [CFU] for bacteria). Mixing data from different units will most likely result in meaningless information – but if we don't know we are mixing units because we didn't explore our data, we are then making decisions based on meaningless information.

- Typos: Often data are input by hand from a device, monitoring probe, or field notebook. As a result, typos are not uncommon. A common example is forgetting to add a negative sign (-) to latitude/longitude coordinates – in which case points plotted in a GIS will show up in the middle of the ocean, or otherwise very distant location. It is helpful to sort your data table (ascending or descending) to see if there are typos in your data (e.g., negative values show up at the top or bottom; numbers with decimal in the wrong place will appear at top or bottom)

- General Trends: simply plotting the data to explore basic trends often helps guide the process of data analysis. Plotting what you measured over time may give a sense for change, and in which direction (increase or decrease). Plotting histograms will help assess if you're data will meet the assumptions needed for families of statistical tests (see *"Musts" for Statistical Tests – Assumptions*)

Error – Type I and Type II

Error is an inherent aspect of all monitoring. It can be introduced by observer bias, equipment used in sampling, or a variety of other reasons. Ultimately, understanding error, it's implications for inferences, and its effect on how confident we are in our results, is a key factor in designing a monitoring protocol. Error is directly related to the question: "is this information better than random?", and thus is a critical piece of monitoring within VSP.

Intuitively, we all are familiar with error. In a previous example we highlighted how different thermometers, with different precision, may give different values for the temperature of boiling water (see *Determining Accuracy and Precision*). The more precise thermometer requires fewer temperature readings to have confidence in the results compared with the less precise thermometer. This is because there is less noise in our monitoring information when we use the more precise thermometer. If we measure the temperature of two boiling pots of water, one pot using the high-precision thermometer and the other the low-precision thermometer, we will likely get different values of temperature between the two pots.

There are generally two types of "error" in monitoring, Type I and Type II error (Tbl. 3). Type I error refers to situations when the null hypothesis (e.g., "no increase in functions and values") is rejected when it is in fact true. A Type II error refers to situations where the null hypothesis is false but is accepted as true.

Using our turbidity scenario as an example, understanding Type I and Type II error looks something like this:

- Overarching question –
    - "Is surface water turbidity increasing over time?"

- Null Hypothesis (i.e., baseline we are testing against) –
    - "There *has not* been an increase in turbidity" since 2011.

- Alternative Hypothesis–
    - "There *has* been an increase in turbidity" since 2011.

Table 3. Possible outcomes from a hypothesis test. Accepting the null hypothesis when it is true, or rejecting when it is false are accurate reflections of conditions. Rejecting the null hypothesis when it is true or accepting the null hypothesis when it is false are inaccurate reflections of conditions.

	Accept Null Hypothesis	Reject Null Hypothesis
Null Hypothesis is True	Correct Decision	Type I Error
Null Hypothesis is False	Type II Error	Correct Decision

You can see from the table that there are only four possible outcomes. If the null hypothesis is in reality true – in this case, that surface water turbidity *is* increasing over time – then your analysis will either conclude that you should accept the null hypothesis (correct) or reject the null hypothesis (Type I error).

Alternatively, if the null hypothesis is in reality false – in this case, that surface water turbidity is *not* increasing over time – then your analysis will either conclude that you should reject the null hypothesis (correct) or accept the null hypothesis (Type II error).

In other words – there is a chance that we will incorrectly accept the null hypothesis even if it's not really true, or, we will incorrectly reject the null hypothesis when it is in fact true. These are Type I and Type II errors. Understanding these errors is essential to designing monitoring protocols that can conclusively answer your question.

Table 4. Possible outcomes from a hypothesis test applied to surface water turbidity.

	Accept (null hypothesis): "there has not been an increase in turbidity"	Reject (null hypothesis): "there has not been an increase in turbidity"
True (reality): no increase in turbidity	(a) Correct Decision	(b) Wrong (Type I Error)
False (reality): no increase in turbidity	(c) Wrong (Type II Error)	(d) Correct Decision

Expanding upon our turbidity example, note Table 4. We would like to know if there has been an increase in turbidity over time. Our null hypothesis is that there *has not* been an increase in turbidity since 2011. If we can conclusively (i.e., based on statistical tests using our data) accept this null hypothesis, then our alternative hypothesis can be rejected. The alternative hypothesis is that there *has* been an increase in turbidity since 2011. If we accept the null hypothesis of "…no increase in turbidity…" based on our statistical test and data, when in reality there has been no increase in turbidity (box *a* in Table 4), then we have made the *correct* conclusion.

If, on the other hand, we reject the null hypothesis of "…no increase in turbidity…", even though it is in reality *true*, then we have committed Type-I error (box *b* in Table 4). That is, there is a mismatch between reality (turbidity *is not* increasing) and our statistical inference (rejecting the null hypothesis […turbidity *is not* increasing…] and accepting the alternative hypothesis […turbidity *is* increasing…]). If we falsely reject the null hypothesis then we may come to the conclusion that turbidity is increasing, when it is not. This mismatch can be profoundly important if you are making decisions based on data you are analyzing; efforts to prioritize the management of turbidity in this (hypothetical) sample area may not be warranted relative to other high-priority areas.

Similarly, if in reality there *is* an increase in turbidity, but our statistical inference leads us to wrongly accept the null hypothesis that there is no increase in turbidity (box *c* in Table 4), again we come to the wrong conclusion. A primary aim of robust design of sampling and monitoring protocols is to reduce the chance that we will inadvertently come to the wrong conclusion based on our information.

How then can we reduce our chances of making erroneous conclusions from sampling and monitoring data[24]? It is helpful to revisit some foundational questions:

(i) What is the "do something level"?

(ii) Is the "do something level" a large change or a small change (e.g., a 5% increase vs. a 200% increase)

(iii) Are you using monitoring methods and data with sufficient signal-to-noise ratios to detect changes? – in other words, looking for a very small change (e.g., a 5% increase) may require more accurate/precise methods and more samples compared with looking for a large increase (e.g., a 200% increase). High signal-to-noise ratio indicates that the monitoring is precise (only a little noise in the data) relative to the "signal" we are looking for. If we are looking for a small change in signal, too much noise (variability) can overwhelm and obscure our results.

Ultimately, errors related to falsely rejecting the null hypothesis (Type-I error) can be minimized by utilizing a smaller P-value to confirm results (e.g., $P < 0.05$ will result in fewer Type-I errors than $P < 0.10$). In other words, setting the limit for what counts as *better than random* to more improbable (5-in-100 chance [P < 0.05] vs. 10-in-100 chance [P < 0.10]) means you are less likely to mistakenly reject the null hypothesis. It is still convention for many environmental science journals to use a P-value cutoff of P < 0.05 for statistical analyses – though there are alternative perspectives on the hypothesis testing, P-values, and the effectiveness of these approaches for understanding the natural world[25] (Murtaugh 2014, Halsey 2019).

Errors related to falsely accepting the null hypothesis (Type-II error) can be minimized by: (i) increasing the sample size of the monitoring approach (e.g., sampling 30 times/year instead of 15 times/year), (ii) using a monitoring approach that is more accurate/precise (i.e., reduces the noise [standard deviation] in the data), and/or (iii) testing for larger differences (e.g., a larger “do something level”).

Thankfully, there are quantitative approaches to estimating how likely a sampling protocol will be able to successful tease out differences – so that watershed work groups can evaluate options for sampling (frequency needed, accuracy needed, etc) for a specific do-something-level and/or for a specific indictor, function, or value of interest. This is called “power analysis” (see *Determining how many samples you need to answer your question: Power analysis*) and is determined in part by tradeoffs between sample size (e.g., 10 samples/year vs. 30 samples/year), noise (e.g., error from sampling approaches), and level of significance used to confirm results (e.g., P-values of 0.01 vs. 0.05 vs. 0.10).

Setting up for success: ensuring your monitoring program is rigorous and meets necessary assumptions – sample size and normality.

The Importance of Adequate Sample Sizes

Small sample sizes often result in undesirable data outcomes. Insufficient sampling results in sample sizes that are too small to capture the signal of interest properly. Moreover, smaller samples are: (i) more likely to result in data that doesn’t fit the normal distribution (and therefore violates assumptions for parametric statistical tests), (ii) more likely to result in inconclusive results (i.e., “…we can’t tell from the data…”), and/or (iii) more likely to result in Type-II errors (accepting the null hypothesis when it should be rejected; see *Error – Type I and Type II*). Ideally, the number of samples collected for monitoring should reflect the goals of the monitoring. In light of these goals: How good of an answer do you need? What is your do-something-level? Does your monitoring approach allow you to sufficiently detect the signal through the noise (i.e., adequate accuracy and precision)? Have you conducted a power analysis? If you can answer the above questions in light of your goals, then you will likely have a sufficiently large sample size. A general rule of thumb is to obtain at least 30 samples for each group/treatment/subject of interest (i.e., if comparing two locations, 60 samples total (30 samples per location).

"Musts" for Statistical Tests – Assumptions

While statistics often seem like "magic" – they are in fact based on mathematical truths and rooted in probability. Essentially, the statistical tests we perform answer questions like: "is this information better than random?", "how much confidence do we have in our conclusions?" In order for the tests to give valid results, we must be sure we are using the right tests and the right data.

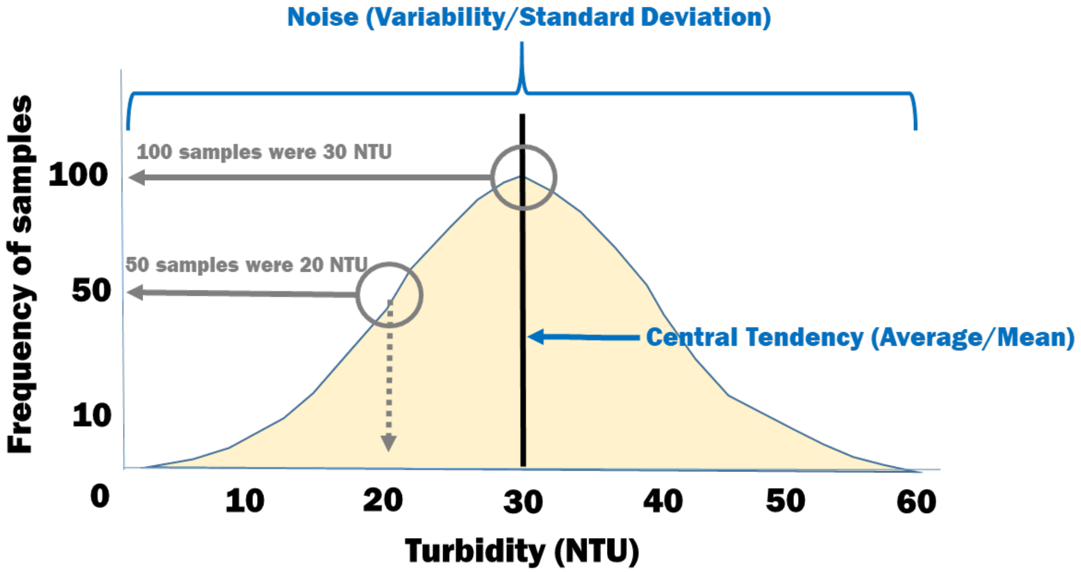
As the old adage goes, "garbage in equals garbage out", and this is particularly true with data analysis and statistics. For example, a diesel engine and a gasoline engine are both "engines" – and each can propel a vehicle down a road under the right conditions. However, if we put diesel in the gasoline engine or *vice versa* our engines will not perform the way we expect them to. The same can be said for statistics. If we put the wrong kind of information in our statistical tests we will come to erroneous conclusions. Fortunately, there a variety of approaches available for different scenarios we may encounter.

Every statistical test that is used is only valid under certain conditions. Some approaches require an equal number of samples between locations/times, some require a linear relationship between the things being measured, and others require that the data follow a "bell shaped curve" (i.e., normally distributed). A partial list of essential assumptions is provided below. A concise summary of statistical assumptions and applications to data is provided in Zuur et al. (2010).

Assumptions for Parametric Statistics

<u>Normality</u> – normality refers to how well data fits the "bell shaped curve". When data follow this shape it is considered to follow a "normal distribution" (also called a "Gaussian" distribution). We can actually get a good sense of the central tendency and noise of our data by looking to see if it follows the normal distribution.

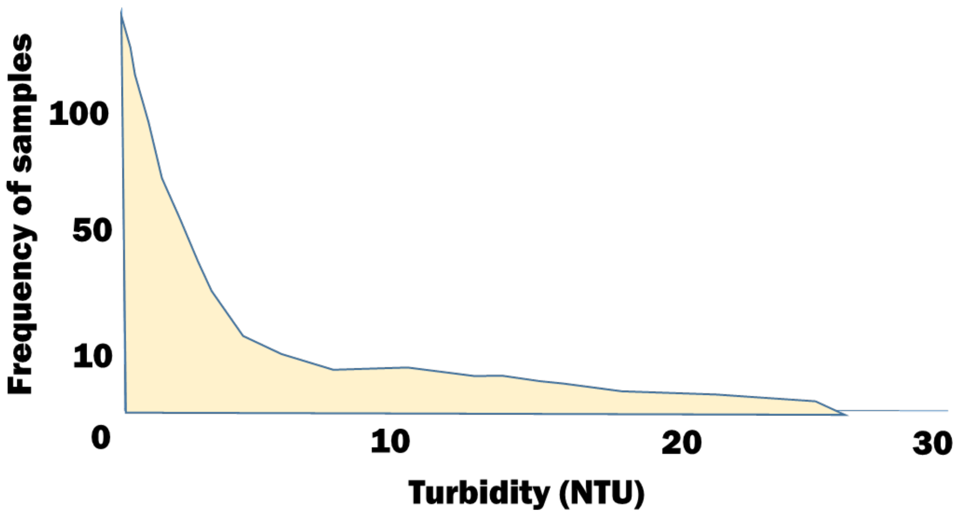
Figure 13. Example of a basic histogram for normally distributed data. The spread of the histogram shows how noisy the data is (variability), the peak shows the central tendency, and the shape (bell-shaped) indicates the data is *normally distributed*.



Note the illustration above (Fig. 13). Along the x-axis are the values of turbidity that were sampled. They range from ~0 NTU to ~60 NTU. The y-axis is the number of times a sample was within a certain value of NTU. For example, about 50 of the samples had a turbidity value of 20 NTU, and about 100 of the samples had a turbidity value of 30 NTU. When taking a turbidity sample, more samples were near 30 NTU than 20 NTU. If any data looks like this bell shaped curve, then the *central tendency* of the data is that value at the peak of the bell. In our example, this is 30 NTU. The *noise* is represented by the width of the bell shaped curve. This represents how far away samples were from the central tendency. An imprecise monitoring method, such as the cheap thermometer referenced in Figure X, will produce wider bell shaped curves than a more precise thermometer.

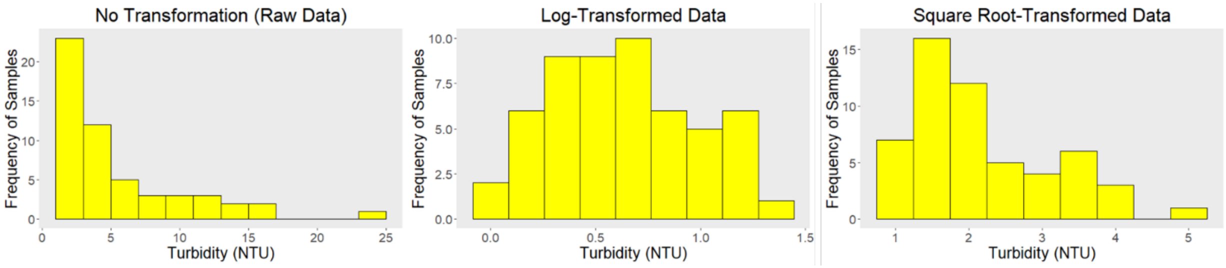
Often, data we collect in the real world does not follow the normal distribution. Consider Fig. 14. The most frequently encountered turbidity level in the samples is zero, with a rather steep decline in frequency until ~5 NTU, where there the turbidity values tail off until the highest value of ~ 27 NTU. Inspecting this histogram, it's not readily apparent what the central tendency is. While the most frequently encountered sample was at 0 NTU, clearly the average level of turbidity is *not* 0 NTU, given that some values were as high as 25 NTU.

Figure 14. Example of a histogram showing turbidity data is *not* normally distributed. The peak of the histogram is ~ 0 NTU, indicating the most commonly sampled value of turbidity was 0 NTU. Most of the samples collected were < 5 NTU with a smaller frequency of values at higher levels, indicated by the tail moving along the right of the x-axis.



The histogram in Fig. 14 is not normal because it does not conform to the bell-shaped curve. As a result, it does not meet a key assumption needed for standard parametric statistical tests. What can we do? First, we should be sure that the data has been explored for outliers or errors – so we can rule out the influence of these factors (see *Data Exploration*). Next, consider applying a data transformation to see if that results in an approximately normal distribution (Fig. 15). If these techniques do not result in the assumptions of parametric tests being met, then consider a non-parametric equivalent (see *Non-Parametric Approaches*).

Figure 15. The left panel is a histogram of the raw turbidity data with no transformations applied. The center panel shows the same turbidity data after applying a log-transformation. The right panel shows the same turbidity data after applying a square-root transformation. Note that the center panel appears more bell-shaped than the raw data in the left panel – this suggests that applying a log-transformation may facilitate the use of a standard parametric test in this case.



<ins>Data are a random sample of the population</ins>: every subject of your target population should have an equal chance of being selected for the sample.

Typically the goal of monitoring is to infer from samples (a subset of a population) some general pattern (the general behavior/condition of the population) that is of interest. For example, a work group may sample 10% of a wetland for plant diversity (a subset of 1/10th of the entire wetland) in order to infer the plant diversity of the entire wetland. In most cases it is not practical or possible to sample 100% of the population of interest. What is desirable is to choose the 10% of the wetland to be sampled in a way that reflects the general condition of the entire wetland. Randomly choosing locations of the wetland to sample until 10% coverage is reached is a way to reduce bias in the results. If the locations of wetland plant diversity samples are left to the field technician to determine based on preference, it is possible that some technicians will sample locations that are easier to access. Easier-to-access locations within the wetland may contain different species composition and abundances than hard-to-access locations. If only the easy-to-access locations are used to infer the condition of the entire wetland, the results will be biased. To avoid such issues, it is recommended to use a random number generator or random number table to choose where to locate samples, so that each area of the wetland has an equal chance of being selected for a sample. The same approach can be used for other functions and values, e.g., choosing areas of a stream reach to sample, or choosing which days to sample a stream reach. Randomizing locations/times of samples can also be done between groups or treatments, which is known as stratified random sampling[26].

Equal "noise" (homogeneity of variance): groups/subjects/treatments have similar levels of "noise".

In other words, we must ensure that our treatment groups (e.g., pre- and post-restoration or upstream and downstream paired samples) contribute roughly the same variability. If for some reason, such as observer bias or equipment calibration issues, variances are very different between the groups/subjects/treatments you are comparing, the results of associated statistical tests are likely to be flawed. Multiple approaches to assess this assumption have been developed, including the F-test, Bartlett's test, or Levene's test[27]. However, simple visualization for comparisons may be sufficient in some cases (Zuur et al. 2010).

Independence of samples: samples are considered independent from each other.

Consider Tobler's first law of geography: "…everything is related to everything else, but near things are more related than distant things…" (Tobler 1970). This applies to relationships in space and in time. In the context of watershed-scale monitoring, if turbidity is sampled in two locations that are only 5m apart from each other, these locations are likely not considered independent – because the conditions affecting turbidity at the two locations are likely very similar. Values of turbidity between these two 5-m-apart samples will likely be more similar to each other than two 5-km-apart samples. Utilizing information from both samples as "independent samples" would result in pseudo-replication – in other words, "false" replication, because we are not getting a lot of new information from a very adjacent location, and the nearby sites are really not independent replicates. It is possible to effectively leverage information from nearby sample locations through the use of advanced techniques, e.g., mixed effects models (groups of nearby samples can be assigned to the same random effect) or utilizing spatially explicit statistical models that estimate and account for spatially correlated data[28] (Lichstein et al. 2002, Zuur et al. 2009). In practice, simple monitoring designs that attempt to limit violations of spatial independence are desirable and achievable.

Tobler's law can also be applied to time – that is, samples taken more closely in time are more likely to be similar than samples taken over longer time periods. Samples of stream flow taken within 5-min of each other will be very similar while stream flows taken within 5-weeks of each other are likely to be more different. There are different ways to factor in temporal correlations[29,30] – e.g., by sampling randomly from a group of observations (e.g., selecting randomly one streamflow measurement per day out of a collection of hourly measurements) or by explicitly accounting for temporal correlations in a statistical model (e.g., adding auto-correlation structure to statistical models)[31] (Zuur et al. 2009). Likely the simplest approach to ensuring temporal independence is to address potential correlation issues before sampling, by developing a robust monitoring plan.

Parametric approaches

The term “parametric” refers to parameters (signal/noise; mean/standard deviation) of a distribution (e.g., of a normal, bell-shaped, histogram of monitoring data). We do not need to get into the weeds on this concept – except to communicate that there is a dividing line in statistics between parametric and non-parametric approaches[32,33]. The most familiar statistical approaches are parametric, these include:

- Paired *t*-test – to compare signals (i.e., means) of the same response (i.e., nutrients, no. of fish, temperature) at the same location/category at two different times to see if they are the same or different[34].
    - Example: Compare average stream temperature at the same location before and after riparian restoration.

- Two-sample *t*-test (a.k.a. Student’s *t*-test) – to compare signals (i.e., means) of the same response (i.e., nutrients, no. of fish, temperature) from two different (independent) locations/categories to see if they are the same of different.
    - Example: Compare average stream temperature between samples of two different streams.

- Analysis of Variance (ANOVA) – to compare signals (i.e., means) between three or more locations/categories to see if they are the same or different (i.e., similar to the two-sample *t*-test but for situations with more than two samples).
    - Example: Compare average stream temperature between samples of five different streams.

- Regression – to determine if the signal of interest (i.e., nutrients, no. of fish, temperature) is associated with another variable of interest (i.e., stream flow, steepness, time)
    - Example: Explore an association between wetland biodiversity and turbidity – e.g., does wetland biodiversity decrease when turbidity increases?

These tests were developed based on data that conforms to a bell shaped curve. The mathematical properties of this curve (also called a “distribution”, e.g., a probability distribution) include terms for the signal (mean) and noise (standard deviation) – and these are essential to ensuring that the tests work right. Not all data will conform to the assumptions required for parametric tests – and if this is the case for your work group’s data, there are options available, such as transforming the data to make it normal (e.g., calculating the base-10 logarithm of values in lieu of original values) or using a non-parametric approach.

Non-parametric approaches

Non-parametric methods do not rely on an underlying probability distribution (e.g., the bell shaped curve), and thus offer some flexibility for situations when data does not conform to parametric assumptions. Non-parametric approaches often utilize a “ranking” approach in lieu of the mean/standard deviation method utilized by parametric statistics. Thus instead of comparing averages between groups, as we do in many parametric approaches, non-parametric methods often compare “average ranks” between groups – so, in other words, with parametric statistics we are not always working directly with the raw data, but often comparing the distribution of ranks between subjects or groups.

Additionally, a pitfall exists utilizing non-parametric statistics when the data already basically conforms to parametric assumptions (e.g., the normal distribution). Generally, non-parametric analysis of a dataset that is approximately normal will have less power to “see” the relationships we are interested in than a parametric counterpart. However, for smaller sample sizes (n < 30), normality of data is often not achieved – in these situations non-parametric statistics may be useful.

A good rule of thumb is to explore your monitoring data to first determine if it (mostly) meets the assumptions required by parametric tests[35]. If so, proceed with parametric tests. If not, see if a data transformation (e.g., log-transforming or square-root transforming) of the data results in a normal distribution. If the data still do not meet the assumptions for parametric tests, then proceed to the non-parametric equivalents[36,37].

Below is a brief list of common non-parametric statistical tests:

- Wilcoxon signed-rank test – to compare signals (i.e., ranks) of the same response (i.e., nutrients, no. of fish, temperature) at the same location/category at two different times to see if they are the same or different.

    - Example: Compare average stream temperature at the same location before and after riparian restoration.

- Wilcoxon rank-sum test – to compare signals (i.e., ranks) of the same response (i.e., nutrients, no. of fish, temperature) from two different (independent) locations/categories to see if they are the same of different.
    - Example: Compare average stream temperature between samples of two different streams.

- Kruskal-Wallis test – to compare signals (i.e., ranks) between three or more locations/categories to see if they are the same or different.
    - Example: Compare average stream temperature between samples of five different streams.

- Spearman's rank correlation – to determine the correlation between two variables of interest.
    - Example: Explore the correlation between wetland biodiversity and turbidity – e.g., does wetland biodiversity decrease when turbidity increases?

Detecting changes over time: what direction are things headed?

Revisiting the *Monitoring Goals* section we can distill common VSP questions into the following:

Is _______ improving/declining/staying the same over time?

This is essentially a "change detection" question because the root of the question(s) involve understanding changes over time. Work groups may wish to know broadly if stakeholder participation, critical area functions and values, or conservation practice implementation are changing over time in order to best manage their programs.

While there are multitudes of approaches to assessing change detection, we will focus here on two approaches: (i) time series analysis, and (ii) grouping by time period. Time series analysis utilizes time as a variable in a statistical model – e.g., using a linear regression with time on the x-axis (Figs. 16 and 17). Practitioners can use the slope of the trend line to infer patterns of change through time. Contrastingly, grouping by time period essentially filters information into "buckets" that refer to specific timeframes (e.g., baseline 2011 vs. a reference time period), and then compares the central tendencies of the buckets to see if they are the same or different (Figs. 18 and 19). These differing approaches utilize data differently, so it is important to consider which method makes the most sense for your group's goals.

Time Series Analysis

At a basic level, time series analysis is fairly straightforward. If the goal is to understand how a phenomena is changing over time, we can simply plot the values of what we have observed or detected on one axis (y-axis) with respect to the time on the other axis (x-axis). Time units can be hours, days, months, years, or decades, depending upon the phenomena being measured.

Figure 16. *Statistically significant* trend lines (solid line) indicate $P < 0.05$. Note that the points are generally more tightly aligned with the trend line when compared to Fig. 17.

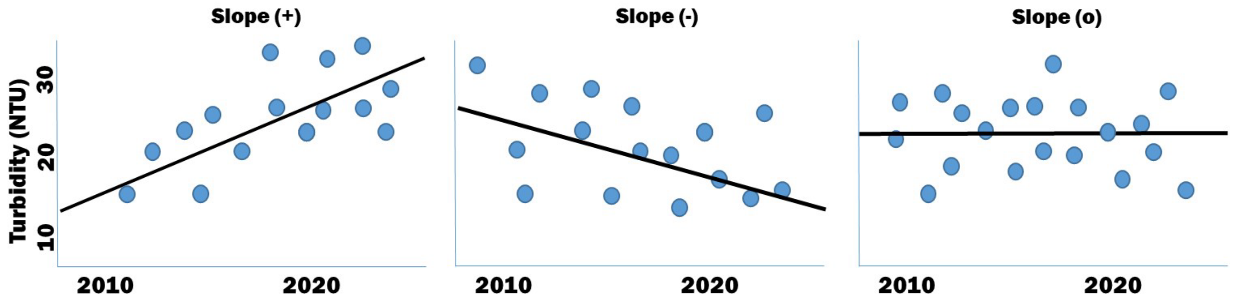
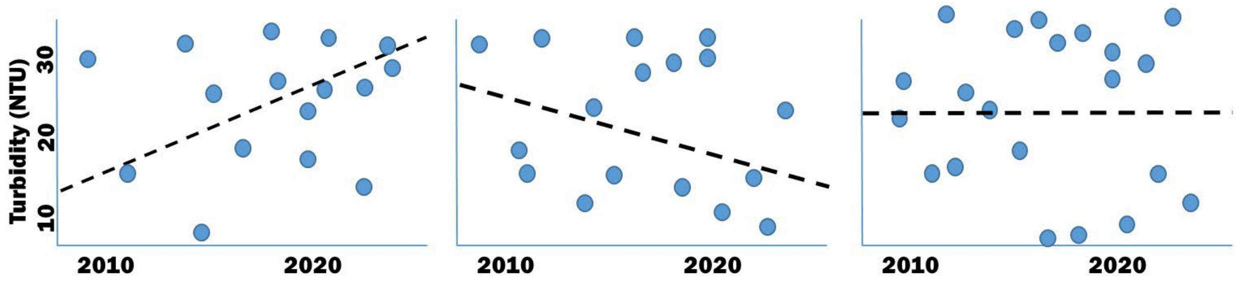


Figure 17. Trend lines (dashed) are *not significant*, indicated by $P > 0.10$. Note that the points are generally more scattered around the trend line when compared to Fig. 16.



With this type of time-series analysis we are interested in knowing whether the slope of the trend line is positive, negative, or zero (no change). We can visually see the pattern from the trend line – but what we wish to know is whether this trend line fits the data in a way that is *better than random*. In a hypothesis framework, our setup might look something like this:

- Overarching question –
    - "Is surface water turbidity changing over time?"

- Null Hypothesis –
    - "There *is no change* in turbidity over time" (slope is zero)

- Alternative Hypothesis–
    - "There *is an increase or a decrease* in turbidity over time (slope is positive or negative) [xxvi]

Note the examples in Fig. 16. The x-axis represents time (years in this example). The y-axis represents the "what" we are tracking (turbidity in this example). On the left side, the slope of the line is positive. In other words, turbidity appears to be increasing. The middle panel indicates turbidity is decreasing, and the panel on the right suggests turbidity is not changing.

However, note that in Fig. 17 the spread of points around the line looks more "noisy" compared to the top. While the trend lines along the bottom panel seem to indicate trends (increase, decrease, or no change), these patterns are not considered *statistically significant* because we don't have high confidence that the data are "better than random". This is indicated by P-values > 0.10. In other words, the trends are *not better than random,* and there is not enough evidence to conclude that the patterns we are seeing are reflective of what's truly happening. While it appears there may be a general trend in one direction or another, we would need better information to have confidence in our conclusions.

For purposes of VSP, we can make better decisions and be better informed if we set up our monitoring in a way that results in trends more like Fig. 16 and less like Fig. 17. We can do this by being clear and concise about what we are monitoring (see *Formulating a*

[xxvi] If you are exploring *any* change (positive or negative slope) then you would use a two-tailed test (two tails for either positive or negative). If you are exploring change in a specific direction only, such as an increase (positive slope), you would use a one-tailed test. More about one-tailed and two-tailed tests can be found in *Volume II: Testing the direction that change is occurring*)

Monitoring Question), how frequently we monitor (see *Sample Timing and Frequency*), and what tools we use to monitor (see *Monitoring Methods*). Better targeted sampling, more accurate/precise sampling approaches, and clearly defined monitoring goals result in more actionable information.

One of the most basic and foundational time-series analysis approaches is linear regression, which is the basis of the examples provided in this section. A linear regression is an equation that predicts an outcome (what you are measuring, e.g., turbidity) in response to changing levels of an explanatory variable[38], e.g., time. However, there are alternatives and derivatives of linear regression that may be more useful depending on the questions of interest and the data available to the work group.

Examples include:

- Non-linear regression – for phenomena that do not behave linearly, e.g., the response variable doubles every month (e.g., population growth)

- Logistic regression – for binary categorical response variables, e.g., yes/no or alive/dead

- Multiple regression – when your response variable is predicted by two or more explanatory variables, e.g., turbidity (response) as a function of stream flow (explanatory variable 1) and slope steepness (explanatory variable 2).

Each of these approaches has their own unique set of assumptions and motivations. More information about the different types of time series analysis can be found in the *Families of Statistical Tests* section.

Grouping By Time Period

Another way of understanding changes through time is to compare average values between two or more time periods. For example, we can group all the data we have collected into two time periods, (i) 2006-2011 (baseline), and (ii) 2012-2017 (comparison) (Fig. 18). We can then compare the mean levels of what we are measuring between the two time periods to determine whether they are different.

The approach to grouping by time period contrasts somewhat with time-series in that we are comparing the central tendencies of two or more time periods to see if there are differences, rather than plotting data over time to estimate the trend (e.g., positive or negative slope of the trend line).

Figure 18. The first panel (a) shows data set up for time-series analysis. The same data is plotted in (b) showing the same data, but grouped into two different time periods, “baseline” and “comparison”.

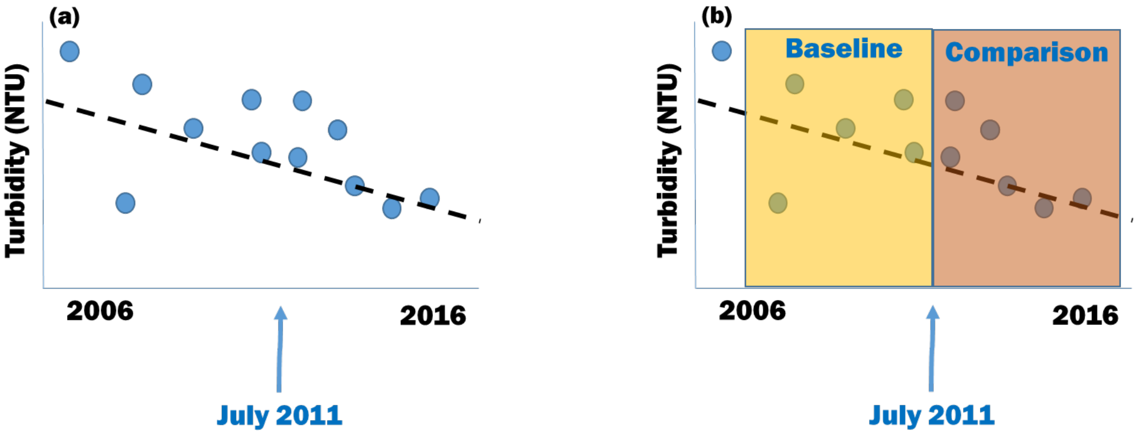
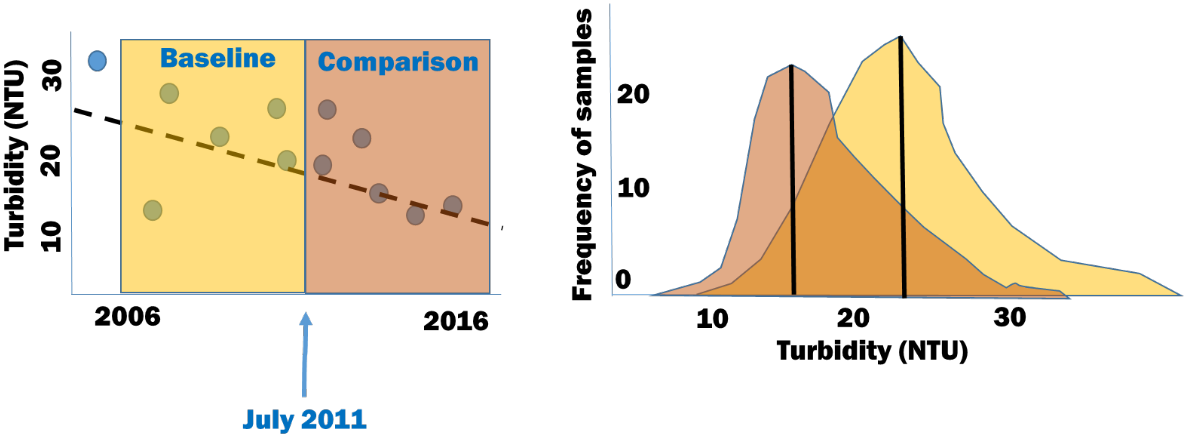


Figure 19. (a) Panel shows turbidity over time, with time periods demarcated by color bands. (b) Panel shows the same data as in *(a)*, but as a histogram with time periods demarcated by color bands (see *“Musts” for Statistical Tests – Assumptions* for more on histograms).

.



To explore this more, note the panels in Fig. 18. The left panel shows data on turbidity tracked over time for roughly a 10-year period. The “baseline” period in this example is a 5-year period between 2006 and 2011, and the “comparison” is a 5-year period between 2011-2016. You can see from the slope of the line in Fig. 18a that turbidity appears to be declining. However, the dotted trend line indicates the trend is not significant (P < 0.10). We can plot histograms of the data from the two time periods, "baseline" and "comparison", and explore the data in a different way, which we have provided in the right panel of Figure 18 (see *"Musts" for Statistical Tests – Assumptions* for more on histograms). The yellow colored histogram corresponds to the "baseline", while the orange color corresponds to the "comparison" period. Turbidity levels are plotted along the x-axis, while the y-axis tells us how many of the samples collected represented a specific turbidity level. The peaks of the histograms are the levels of turbidity that were collected the most – that is, of all the samples collected during the "comparison" period, turbidity levels of ~ 15 NTU were the most frequent; of all of the samples collected during the "baseline" period, turbidity levels of ~ 25 NTU were most frequently collected. You can see this visually by the vertical black lines indicating the peaks of each time period's histogram.

Ultimately, with this type of analysis, we are interested in knowing whether the central tendencies between groups are the same or different. In a hypothesis framework, our setup might look something like this:

- Overarching question –
    - "Is surface water turbidity changing over time?"

- Null Hypothesis –
    - "There *is no difference* in average turbidity between the 'baseline' and 'comparison' time periods

- Alternative Hypothesis–
    - "There *is a difference* in average turbidity between the 'baseline' and 'comparison' time periods.

Like with time-series analysis, the basic visual trends are helpful, but we are aiming for knowing with some certainty if the trends are *better than random*. In the hypothesis framework, we want to know if we have enough evidence to reject the null hypothesis – and this depends on the signal-to-noise ratio, sample size, accuracy/precision of our monitoring data, and our standard for rejecting or accepting the null hypothesis (e.g., $P < 0.10$ vs. $P < 0.001$).

Figure 20. Panel (a) illustrates a trend in turbidity over time, with time on the x-axis and sampled turbidity level on the y-axis. Panel (b) shows the same data as a histogram, with sampled turbidity levels on the x-axis and the frequency of samples collected for a given turbidity level along the y-axis. Data in this example have considerable noise (variability) – observable from the spread of points around the trend line in panel (a) and the relatively wide and short shape of the histograms, and overlapping of histograms between time periods, in panel (b).

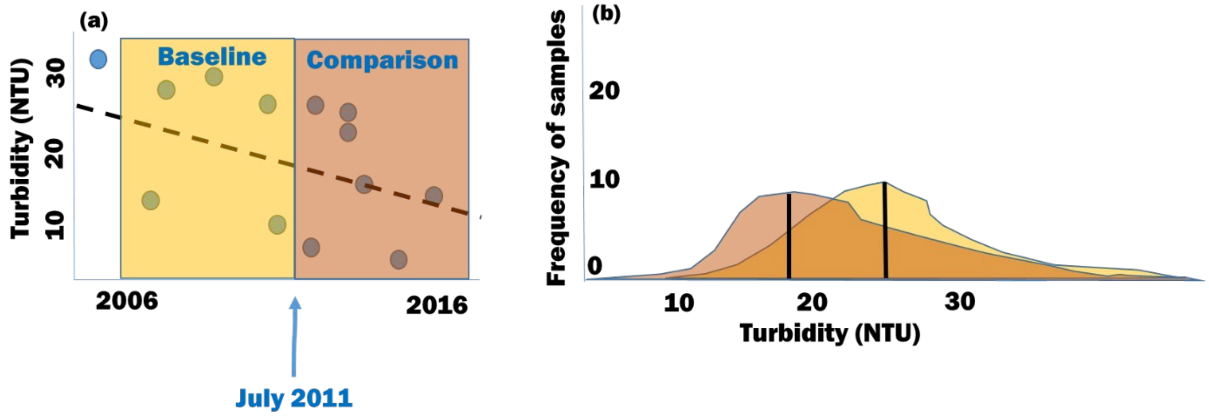
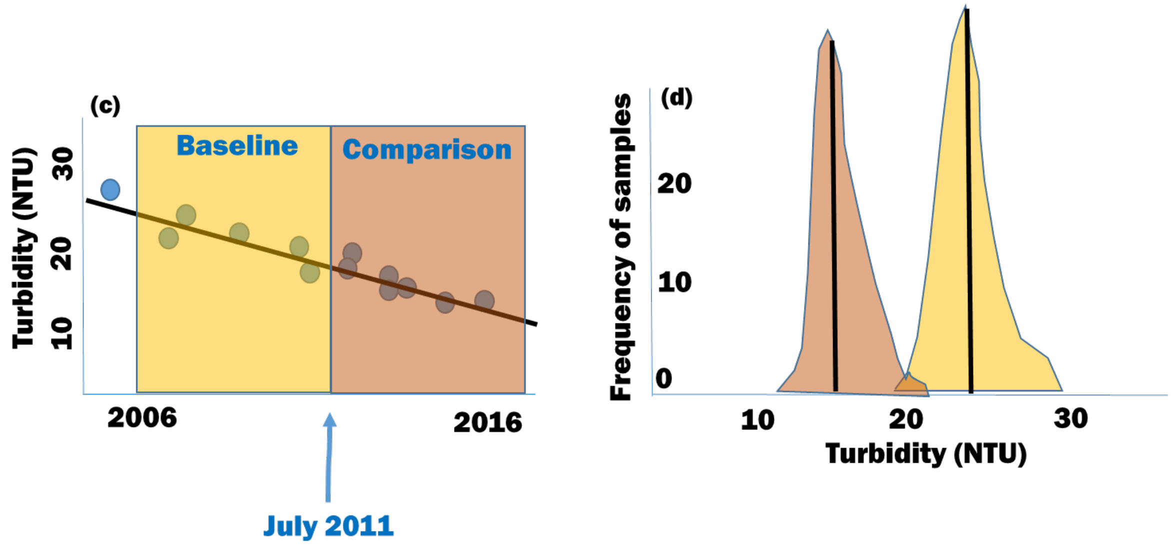


Figure 21. Panel (a) illustrates a trend in turbidity over time, with time on the x-axis and sampled turbidity level on the y-axis. Panel (b) shows the same data as a histogram, with sampled turbidity levels on the x-axis and the frequency of samples collected for a given turbidity level along the y-axis. Data in this example have less noise (variability) compared to Fig. 21 – observable from the tighter spread of points around the trend line in panel (a) and the relatively narrow and tall shape of the histograms, and non-overlapping of histograms between time periods, in panel (b).



Consider Figure 20. Panel (a) shows turbidity levels grouped by time period. The points do not fit the trend line particularly well, indicating some “noise” in the data. Panel (b) shows the same data but as a histogram. Note that the peaks in the histograms are less pronounced than in Fig. 21 and there is also quite a lot of overlap between histograms from the two time periods. This indicates that the central tendencies of turbidity values between the two time periods are *not* different. Contrast this with panels (c) and (d), where the turbidity values are tightly oriented around the trend line and the histograms almost do not overlap at all. This indicates that the central tendencies of turbidity values between the two time periods *are* indeed different.

Thus, the same considerations regarding the robustness of data, monitoring approaches, and question formulation pertain to grouping by time period as they do for time series analysis. The primary difference between the two approaches is the way the data is analyzed, explored, and tested for significance.

Testing the direction that change is occurring: One-tailed vs. two-tailed tests

Because monitoring within VSP is focused on detecting change (or no change) over time, it is important to understand which direction of change you are assessing. Is your watershed work group interested in knowing only if there’s been (i) an increase, (ii) a decrease, or (iii) any type of change (either increase of decrease)?

The scope of interest will dictate whether you use a one-tailed or a two-tailed statistical test. Recall that within a hypothesis framework (see *Volume II: Monitoring Design, Hypotheses*) statistical tests tell us whether we can reject our null hypothesis in favor of the alternative hypothesis and provides a measure confidence we have to make this decision (e.g., a P-value and/or confidence interval). If the alternative hypothesis is *directional*, in that it specifies a direction of change – either an increase only or a decrease only, then you may consider using a one-tailed test. If the alternative hypothesis is *not directional*, and the overall goal is to assess whether or not any change has occurred – such as either an increase or a decrease, then you may consider a two-tailed test.

A summary of various statistical tests can be found in Appendix-2. When comparing only two populations, time periods, or treatments, each of the tests may be conducted as either a 1-tailed or 2-tailed test (when comparing 3 or more groups, such as with ANOVA, the concept of 1- or 2- tailed tests is irrelevant). The term “tail” refers to the tail-like shape on either side of a normal distribution that encompasses the significance level, e.g., $P < 0.05$ (Fig. 19). In a one-tailed test, we are looking to see if the central tendency of the comparison data is below (decrease), or above (increase), the baseline central tendency (Figs. 22*a* and 22*b* respectively). In other words, the direction of the trend to be assessed (above or below) is determined prior to conducting the test, and the results of the test tell you if a significant trend in that direction occurred. A two-tailed test is used if we looking to see if the central tendency of the comparison data is different from the baseline, either above or below the central tendency of this baseline (Fig. 23). Two-tailed tests do not require the direction of the trend prior to the test as with one-tailed tests.

Figure 22. Example of one-tailed tests applied to dissolved oxygen over time relative to baseline conditions (e.g., conditions as of July 2011) using P < 0.05 as the level to reject the null hypothesis. In this example, the comparison value of dissolved oxygen is within the upper tail (*a*) or lower tail (*b*), suggesting we can reject the null hypothesis that there is no difference between the baseline and comparison. *(a)* Comparison central tendency is greater than the baseline central tendency, indicated by the blue region above the baseline average, and representing a < 5% chance of wrongly rejecting the null hypothesis. *(b)* Comparison central tendency is less than the baseline central tendency, indicated by the blue region below the baseline average, and representing a < 5% chance of wrongly rejecting the null hypothesis.

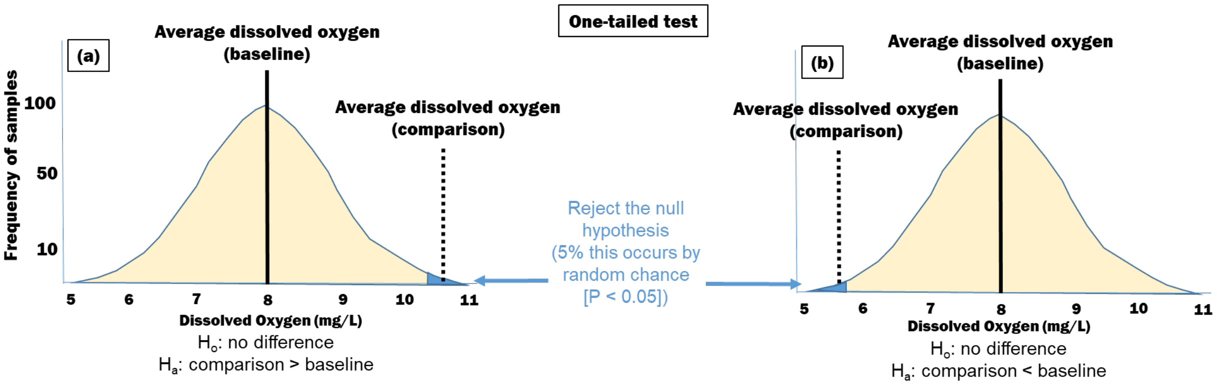
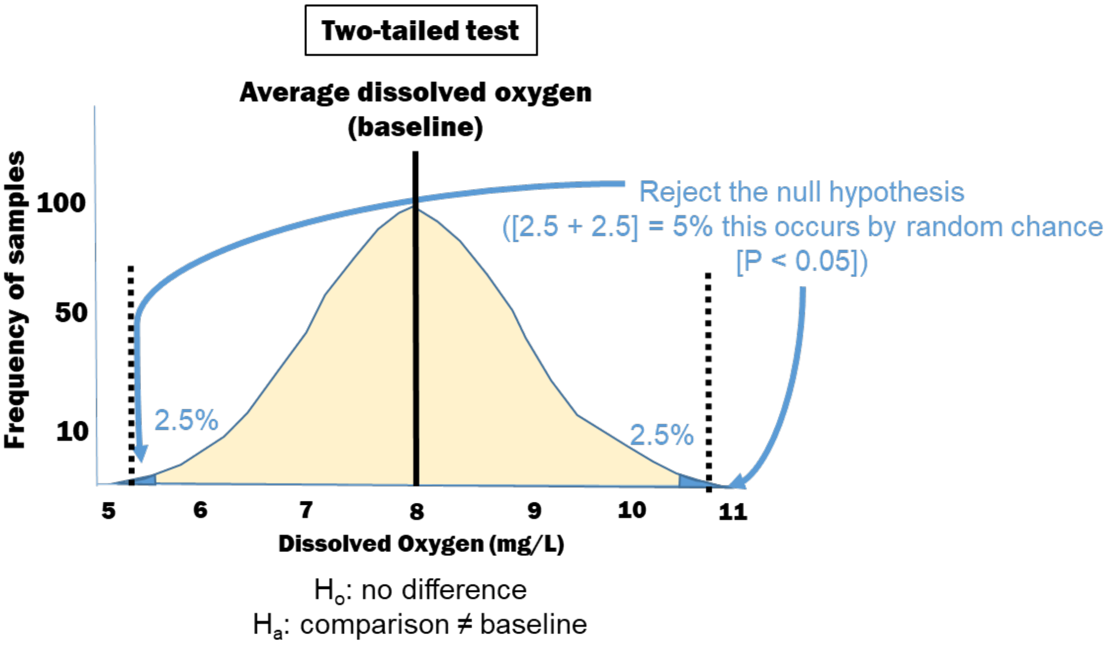


Figure 23. Example of a two-tailed test applied to dissolved oxygen over time relative to baseline conditions (e.g., conditions as of July 2011) using P < 0.05 as the level to reject the null hypothesis. Note that each tail comprises 2.5% of the significance level, i.e., 2.5% x 2 tails = 5%.



In short, if the effect you are interested in can only go one direction, such as an increase only, then you may wish to use a 1-tailed test. The advantage of this approach is that you will have more statistical power to detect an effect in one direction compared with a two-tailed test. The disadvantage of a 1-tailed test is that you may fail to detect a change in the opposite direction if one exists. Using a 2-tailed test can detect an effect in either direction (increase or decrease) but has less statistical power to detect effects overall because the significance level is shared between two tails[39].

Also note that statistical significance of one- and two-tailed tests depends upon the level of significance desired by the work group (e.g., P < 0.05 vs. P < 0.10), the precision and accuracy of the monitoring methods used, and trade-offs with respect to Type I and Type II error (see *Understanding your data: Trends, noise, and errors*). Figs. 22 and 23 refer to approaches with a significance level of P < 0.05.

Determining how many samples you need to answer your question: Power analysis

Power analysis essentially estimates the "power" of a data set to confirm or reject a hypothesis when the goal is to determine whether there are differences between phenomena (e.g., is turbidity higher in stream A than in stream B?). For example, if the hypothesis is "there is no difference in stream turbidity between today and the baseline conditions", then a *power test* can illustrate how many samples you will need (approximately) to be able to show this in a statistically meaningful way. In short, the goal of power analysis is to leverage existing data (such as from a pilot study, freely available data, or previous monitoring that was inconclusive) or estimates of data (based on literature or best available information regarding the signal-to-noise of the phenomena of interest) to determine how many samples are needed to reject or accept a null hypothesis.

In the context of VSP, power analysis in an elegant way to tie together several themes: “What level of precision/accuracy do you need?”, “When will you do something with the monitoring answer?”, “How often are we monitoring?”, “Are the results better than random?”.

Recall that in a hypothesis framework, accepting or rejecting a null hypothesis depends upon the threshold of the level of uncertainty we are willing to accept (see *Volume II: Monitoring design - Hypotheses*). Thus it is critical to ensure that whatever monitoring approach is used has an appropriate accuracy/precision level for the questions identified by the watershed work group (see *Volume I: What level of precision/accuracy do you need?*). Almost always, a key goal in developing hypotheses is to detect an effect or difference of some kind, if one exists, and to have some idea of how much confidence to have in the results – which is how P-values can inform our hypothesis tests. Your monitoring plan should explicitly state the confidence required to confirm or reject a hypothesis (see *Volume II: Monitoring design - Statistical considerations for monitoring within VSP*).

A practical consideration for monitoring is the determination of how many samples need to be collected in order to answer the monitoring question of interest. Power analysis utilizes the inherent noise of the signal, as well as thresholds identified by practitioners for the signal (i.e., the do-something-level), to estimate monitoring outcomes based on the total number of samples to be collected.

For example, a do-something-level for nutrients that is an increase of 200% will require fewer samples than an increase of 5%. This is because it’s easier to “see” very large changes compared to very small changes if the underlying noise of the data is the same. Deciding before-hand what the do-something-level is, and incorporating it into your monitoring plan, is essential to developing a robust monitoring protocol.

Power analysis uses four key ingredients, which are related to each other, to estimate the ability of the data to answer a specific question (i.e., based on a hypothesis) – *sample size*, *significance*, *statistical power*, and *effect size*. *Sample size* refers to the number of observations or samples collected. Generally, larger sample sizes reduce error and improve the ability to "see" changes if they exist. *Significance* refers to the desired significance required to reject/accept a hypothesis (e.g., a 5% [P < 0.05] or 10% [P < 0.10] chance the results occurred by chance [Tbl. 2]). Recall that the significance level chosen relates to how likely the null hypothesis is rejected but is in fact true (Type I error; Tbls. 3 & 4). *Statistical power* refers to the likelihood that the null hypothesis is accepted when it is in fact not true (Type II error; Tbls. 3 & 4). Statistical power of 80% (0.80) suggests that there is a 20% chance of a Type II error. *Effect size* refers to the magnitude of the change in signal we are assessing (e.g., difference in mean values between the baseline and the comparison) relative to the "noise", or variability, in the data. Effect size is conceptually similar to the signal-to-noise ratio – in that it is hard to "see" small changes if there is a lot of noise in the data. Data collected with high precision and accuracy will reduce noise in the data. Collecting a larger number of samples or observations may also reduce variability in the data. Likewise, selecting a larger effect size can reduce the number of samples needed for data with the same variability.

Note that the four key ingredients are highly related. Practitioners must select the values for three of the ingredients and then solve for the fourth ingredient. If the goal is to estimate how many samples or observations are required to effectively answer the questions outlined by the watershed work group, this will typically require choosing the desired significance level, statistical power, and effect size, and then solving for the sample size. Typical values for significance level are 0.05 – 0.10 (Tbl. 2). A common level for statistical power is 0.80. When comparing "signals" to identify change, effect size is often calculated using Cohen's *d*, which is the difference in signal between the baseline and comparison relative to the "noise" of the data *([mean baseline – mean comparison] / [pooled standard deviation of the sample] = effect size*). Effect size is the "do something level" in relation to the inherent noise in the measurements or observations. More details regarding the elements of power analysis, and how to calculate the desired ingredients, are addressed in EPA's Minimum Detectable Change and Power Analysis Technical Memorandum #3[40]. Additionally, an excel workbook with the calculations necessary for power analysis was developed by the EPA and ECY and is available for use[41].

Worked example of implementing a power analysis: water quality from a single surface water station

To better illustrate the concept of power analysis, consider water quality information for one station in Ferry County sampled for dissolved oxygen. Data were retrieved from the

National Water Quality Monitoring Council's Water Quality Portal (WQP)[42]. The WQP contains water-related data from a variety of organizations, projects, and programs. A database of similar water-related information for Washington can be found using ECY's Environmental Information Management System (EIM)[43].

Data were selected from the timeframe 2006 to 2016. Samples collected from 2006 to 2011 were chosen as the "baseline", and data from 2011 to 2016 were chosen as the "comparison". The baseline period contained 29 samples and the comparison contained 26 samples (Fig. 24a). The "signal" indicates an average level of dissolved oxygen of 10.4 mg/L in the baseline period and 9.9 mg/L in the comparison period. Note that the "noise" in the dissolved oxygen data is fairly low, indicated by the relatively narrow standard error lines in Fig. 24a. Data are approximately normal (Fig. 24b).

Fig. 24a. Plot of mean dissolved oxygen concentrations (colored bars) with associated "noise" (standard deviation lines) for one station, for two time periods: baseline (2006-2011) and comparison (2011-2016)

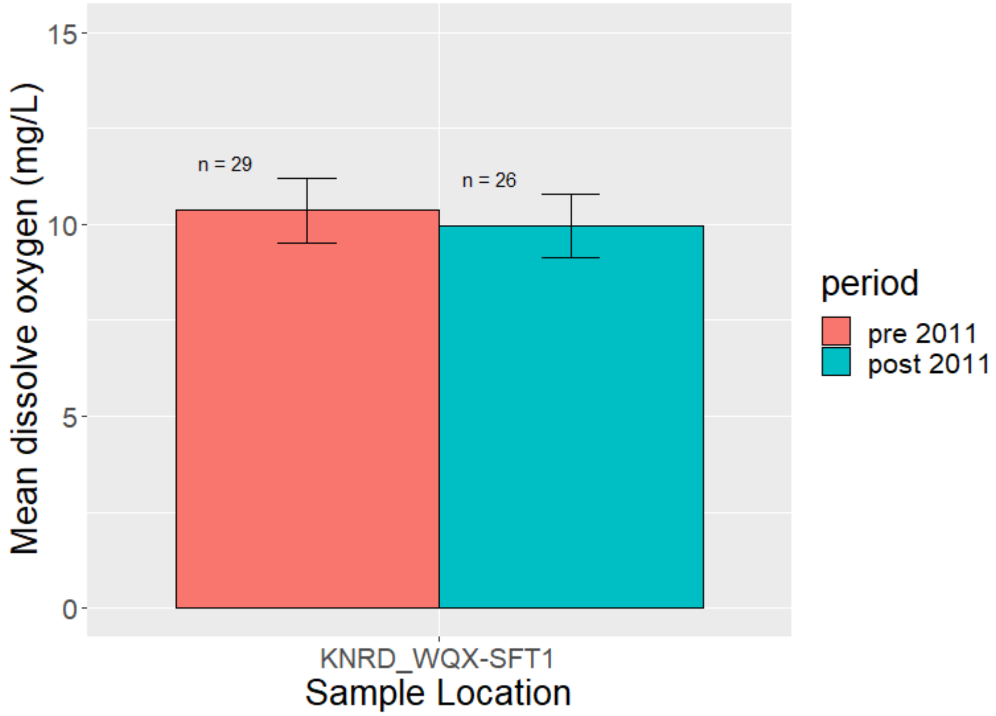
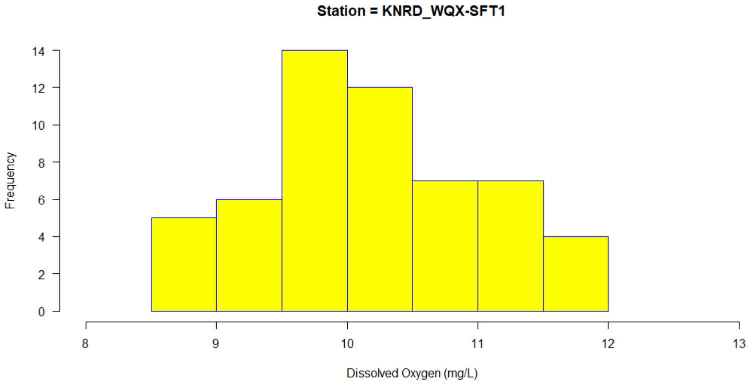


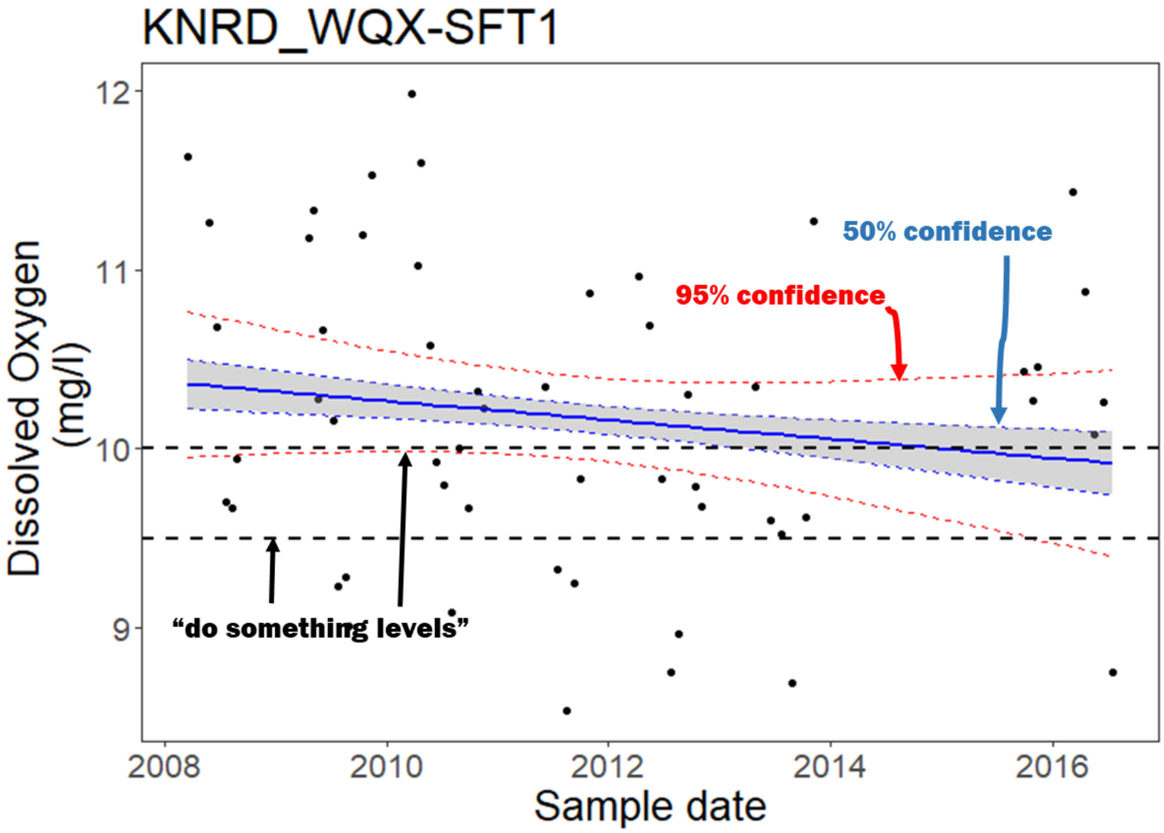
Fig. 24b. Histogram plot of dissolved oxygen exhibiting an approximately normal distribution.



Recall the importance of establishing a "do something level". For purposes of this example, two different do-something-levels for dissolved oxygen are used – 10 mg/L and 9.5 mg/L. Dissolved oxygen criteria for lakes and streams are provided using guidance from ECY[44] (> 9.5 mg/L = *extraordinary*, > 8 mg/L = *excellent*, > 6.5 mg/L = *good*, and > 4 mg/L = *fair*). The do-something-level 9.5 mg/L corresponds with moving from a status of *extraordinary* to *excellent*. The do-something-level of 10 mg/L represents a smaller change and is provided as a contrast.

A basic time-series plot was generated along with each of the do-something-levels for reference. The estimated mean dissolved oxygen is indicated with the thick blue line (Fig. 24c). The dashed red lines on either side of the mean represent 95% confidence levels and the shaded area represents the 50% confidence level. Based on Fig. 24c, we can say we are 50% confident the mean value of dissolved oxygen is greater than the do-something-level of 9.5 mg/L. The 95% confidence interval overlaps only slightly with 9.5 mg/L level – mostly due to high variability (noise) and relatively few samples between 2014 and 2016. But for the most part, there is no overlap between the 95% confidence interval and our 9.5 mg/L do-something-level. Because the do-something-level of 10 mg/L falls within both the 95% and 50% confidence intervals, we cannot have great confidence that the true value of dissolved oxygen is above 10 mg/L.

Fig. 24c. Time series plot showing the trend in dissolved oxygen concentration over time for one surface water station with 50% and 95% confidence levels indicated.



Our null hypothesis is that there is no decrease in dissolved oxygen between the baseline and comparison time periods.

- Overarching question –
  - "Is surface water dissolved oxygen decreasing over time?"

- Null Hypothesis –
  - "There *is no decrease* in dissolved oxygen over time"

- Alternative Hypothesis–
  - "There *is a decrease* in dissolved oxygen over time"

Because we are comparing signals of the same response (i.e., dissolved oxygen) at the same location (station KNRD in Ferry Co.) at two different times (2006-2011 and 2011-

2016) we will use a paired *t*-test for our power analysis (Appendix-2). Table 5 lists the values for significance, effect size, and statistical power used for this example.

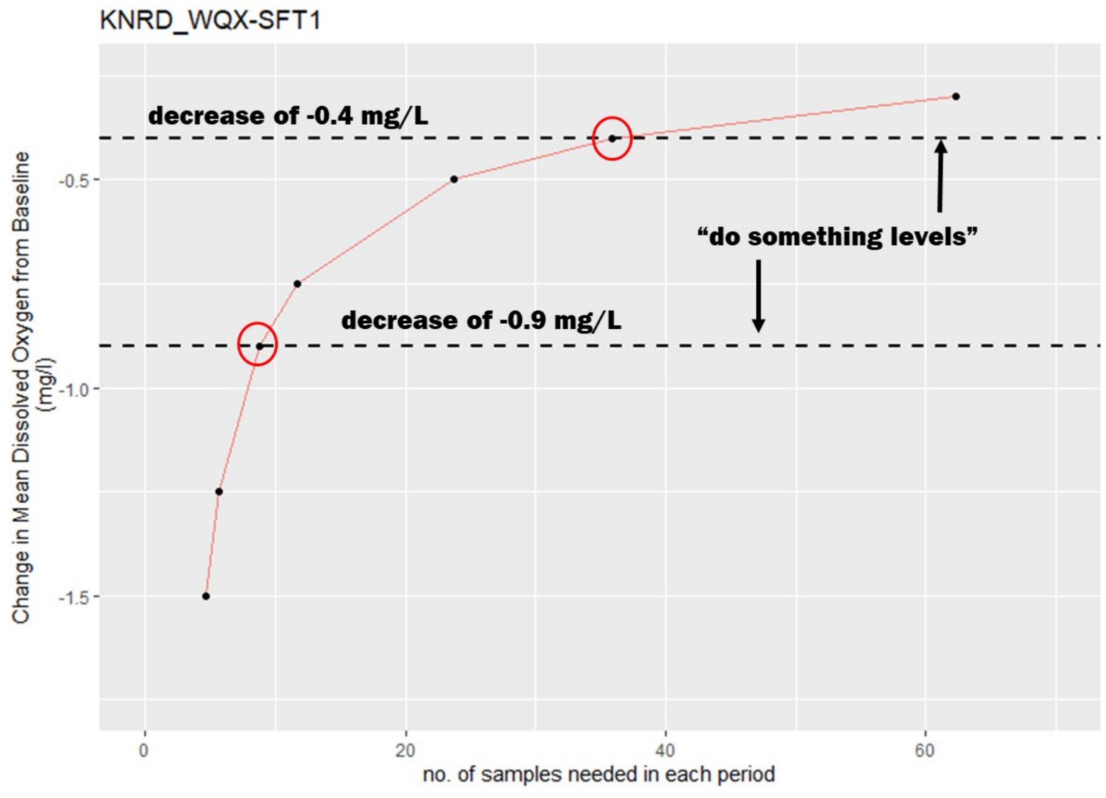
Table 5. Table of parameters used for the power analysis demonstrated in the worked example.

Statistical test	Paired *t*-test
Direction of change	Decrease; one-tailed
Significance level (*alpha*) (*chosen*) (probability of Type I error = 5%)	P < 0.05
Statistical power (*chosen*) [1 - Type II error] = statistical power (probability of Type II error = 20%)	0.80
Pooled variance ("noise" in the data, pooled for both comparison and baseline periods – *calculated from the sample data*)	0.83
Baseline mean dissolved oxygen (*calculated from the sample data, 2006-2011*)	10.4 mg/L
Do-something-level	10 mg/L or 9.5 mg/L
Signal change from baseline to do-something-level of: 10 mg/L	10.4 mg/L – *0.4 mg/L* = 10 mg/L
Signal change from baseline to do-something-level of: 9.5 mg/L	10.4 mg/L – *0.9 mg/L* = 9.5 mg/L
Effect size for do-something-level (Cohen's *d*)	[Signal change from baseline] / [Pooled variance] = [Cohen's *d*]
Cohen's *d* effect size *(for 10 mg/L do-something-level)*	-0.48
Cohen's *d* effect size *(for 9.5 mg/L do-something-level)*	-1.08

Using the criteria listed in Tbl.5 allows for the calculation of the number of samples needed to answer our questions. A graphical illustration is provided in Fig. 24*d*. Each of the two "do something levels" are indicated by the horizontal dashed lines. The number of samples needed (each period – baseline and comparison) to conclusively answer our questions is represented on the x-axis. The change in dissolved oxygen we are interested in is listed along the y-axis. Note that smaller changes in dissolved oxygen require more samples in order to "see" the change in a way that's better than random (i.e., statistical power of 0.8 and significance level of P < 0.05). In order to know if dissolved oxygen levels have decreased to 10 mg/L (a change of -0.4 mg/L from the baseline average of 10.4 mg/L), then we would need approximately 38 samples from the baseline period and 38 samples from the comparison period (total of 76 samples). If we are only interested in a large change, such as a decrease of 1.5 mg/L, then we would only need a handful of samples in each period to demonstrate this.

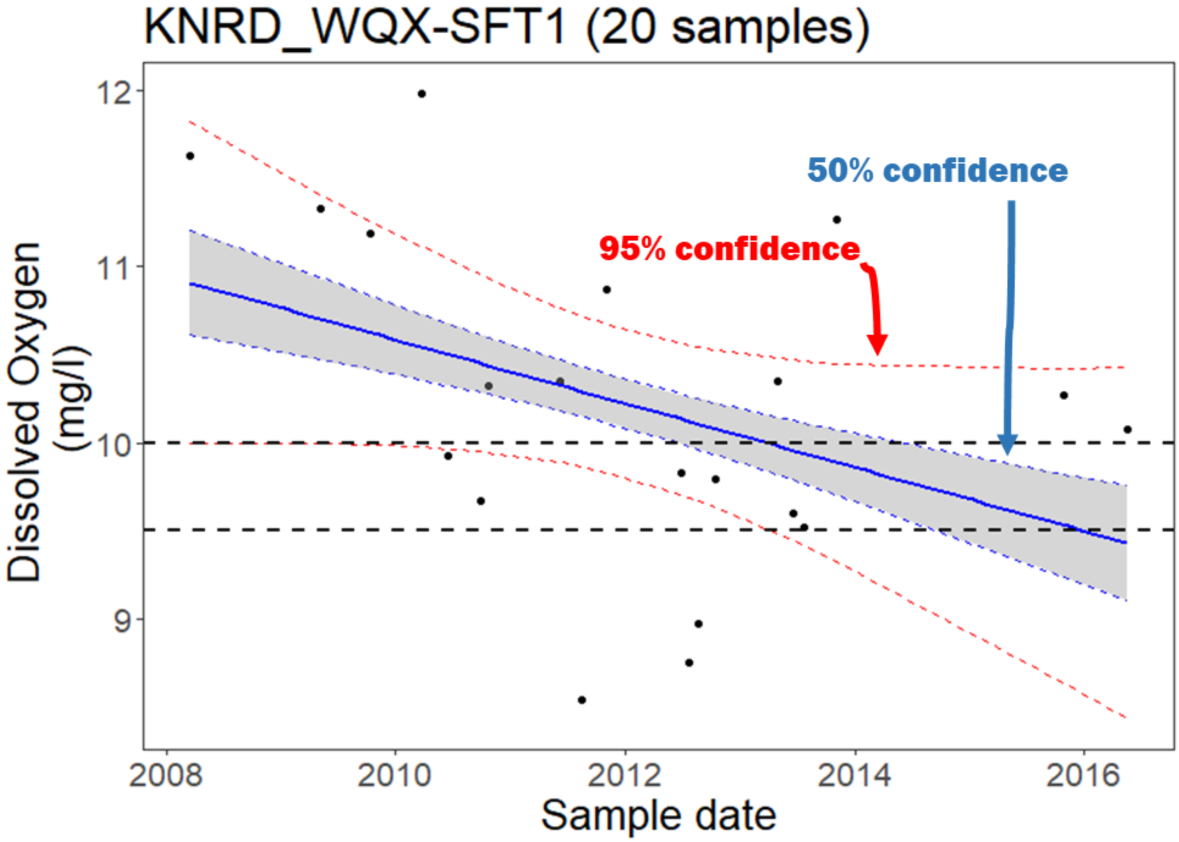
Watershed work groups may be interested in different magnitudes of change for different locations, or functions and values, within a watershed. Approaching monitoring with a “do something level” will help set the stage for establishing a reasonable number of samples or observations to be conducted relative to the goals of the watershed work group. Sample size is a large contributor to the cost of monitoring, so it’s important to balance meaningful threshold detection with efficient sampling to end up with a robust statistical conclusion. It’s also important to remember that these estimates are based on the inherent “noise” of the dissolved oxygen data – if a new method is used that is less precise, there will be more noise. More noise will require more samples to “see” changes relative to your do-something-level.

24d. Results of a power analysis with the number of samples needed (x-axis) to see a specific change from baseline dissolved oxygen concentrations (y-axis). Detecting smaller changes requires more samples.



We can see that based on this approach, the concentration of dissolved oxygen at this monitoring station has likely *not* decreased below 9.5 mg/L (Figs. 24*a* & 24*c*). However, imagine that we only had 20 samples total instead of the original 55 samples (29 samples for the baseline and 26 samples for the comparison). Smaller sample sizes result in less confidence in results. A plot of dissolved oxygen is presented in Fig. 24*e* in which a random subset of 20 samples from the original data was used. Note that the decrease looks more pronounced compared to the full data set in Fig. 24*c*. The 50% and 95% confidence intervals are wider, indicating less certainty in the results. ***If we had only sampled this location 20 times, we would likely not be able to answer our question conclusively.*** Moreover, with only 20 samples, it appears that there is a more dramatic decline in dissolved oxygen, when this is not in fact occurring (Fig. 24*c*).

24e. Time series plot showing the trend in dissolved oxygen concentration over time for a subset of only 20 samples from the original data (Fig. 24*c*).



In summary, power analysis ties together several important monitoring themes. The precision of your monitoring approach affects the amount of “noise” in the measurements taken, and therefore dictate how many samples are needed to conclusively determine whether change has occurred. Additionally, some themes of monitoring are inherently more noise than others – e.g., bacteria concentrations tend to be noisier than dissolved oxygen concentrations in many systems. This type of context may be identified from the various monitoring projects and data cited in many of the VSP watershed work plans.

Establishing a “do something level” provides a target to estimate if current sampling is sufficient or if more, or less, sampling is needed. The desired statistical significance level and statistical power (which relate to acceptable levels of error; Tbl. 5), chosen by practitioners informs what is considered “better than random” with respect to results. Approaching monitoring with these principals in mind will facilitate efficient, effective, and actionable monitoring.

Other considerations: monitoring boundaries, timing, and frequency

Monitoring boundaries

It may be useful to revisit the purpose of monitoring within VSP (see *Volume I: Formulating a Monitoring Question*) to consider what the “boundaries” of your monitoring program are. Within VSP, it is not practical or necessary to monitor *everything* that is occurring within the watersheds of your county. Recall that “monitoring” refers to watching or tracking something over time for a special purpose. If the watershed work group is clear on the *special purpose* of the monitoring (from language in the watershed work plan), it will be easier to identify the relevant boundaries of your monitoring program. By understanding and defining these boundaries a watershed work group is more likely to be able to simplify and target monitoring to better measure the “signal” of agricultural activities most relevant to VSP. The purpose of this section is to facilitate narrowing the scope of monitoring so that it is practical and well-suited to the questions your watershed work group intend to address.

Clarifying the boundaries of your monitoring plan will help target sampling to hone in on the signal of interest and minimize the effects of non-VSP-related influences, such as climate change, industrial pollution, or non-agricultural-related wildfire. Targeted sampling, tailored to the specific questions outlined by your watershed work group (see *Volume II: Monitoring Design – Monitoring goals*) is likely to provide more actionable information.

In order to establish the boundaries of your monitoring program it may be necessary to revisit the contextual information relating to: (i) the natural resource issues of most concern, (ii) the historic (antecedent) conditions of agriculture and the ecosystems within the county’s watersheds, and (iii) the regional influences or activities outside the control of the county (e.g., neighboring counties, upstream watersheds, connected groundwater systems, regional changes in precipitation or temperature patterns, or ocean conditions affecting salmon). *Much of this information is contained within your county’s watershed work plan or sources cited therein*.

For example, some stream systems may have sediment contributions from a suite of drivers, such as stream channel erosion, urban development, and agricultural practices. Urban development on non-agricultural lands within an urban growth boundary is not considered an agricultural activity that falls within VSP – but measurements of stream turbidity may still detect a signal from this type of development depending on where and when they are collected (Fig. 25). Therefore it is useful to know how “background” conditions are contributing to what it is that you are sampling or measuring. For more information on this type of contextual information see *Volume I: How local and regional conditions affect monitoring programs*.

Particularly for surface water, influences and activities occurring upstream of any county will undoubtedly affect samples or measurements within the county. Having some idea – or estimate – of how contaminant concentrations, invasive species, etc, are flowing into your county from a neighboring area will aid in disentangling agricultural effects from other sources (Fig. 26).

Figure 25. Conceptual illustration of soil erosion occurring from agricultural activities and urban development. In surface water, the levels of sediment found in water depends upon activities upstream of the sampling location. The signal of agriculturally-related soil erosion will be most clear if the site *ag* is sampled. The signal of agriculturally-related soil erosion will be absent if site *urb* is sampled, as there are no agricultural activities in the associated sub-watershed. Samples of sediment in water collected at the site *both* will include signals of erosion from both urban development and agricultural activities.

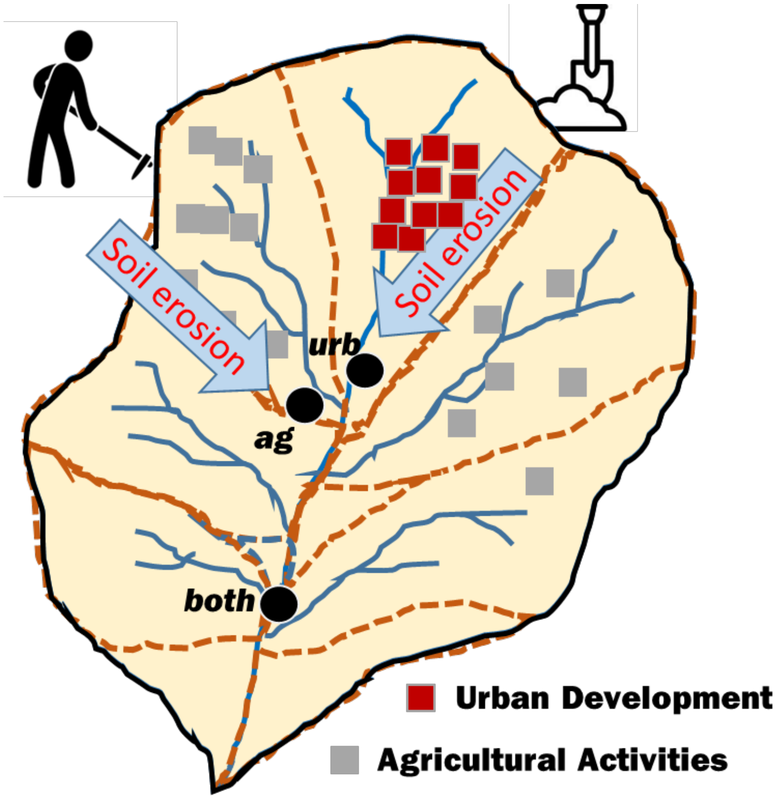
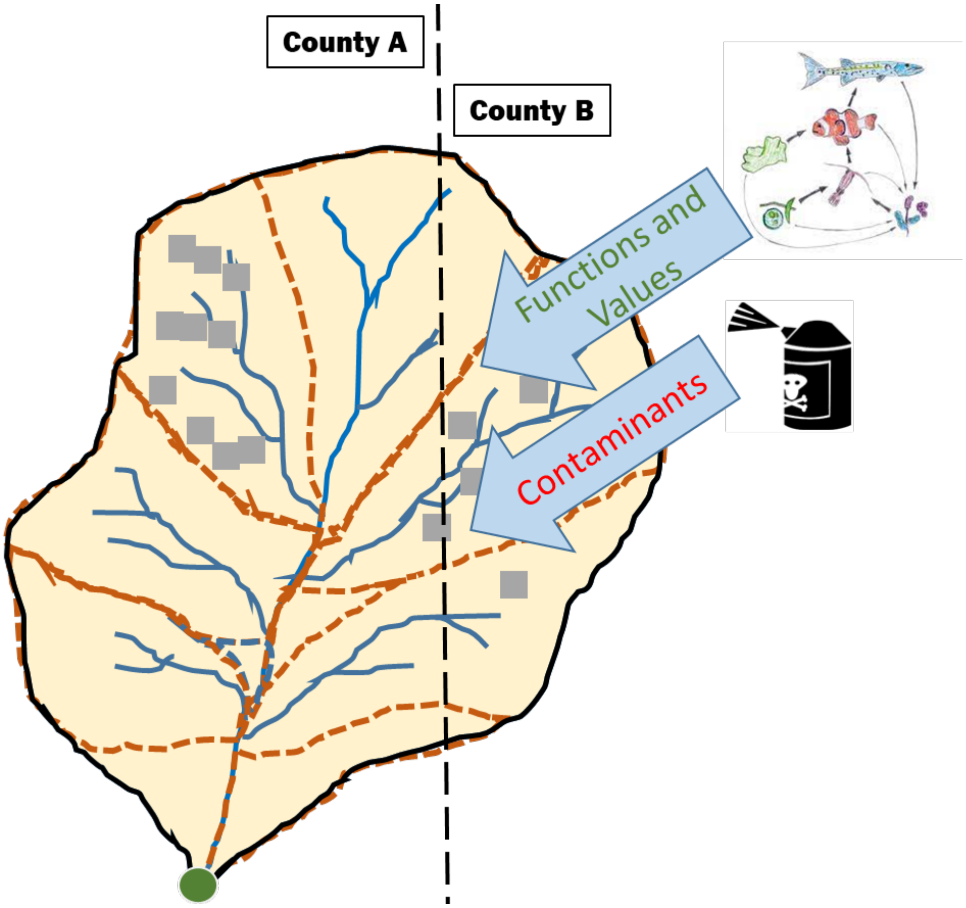


Figure 26. Illustration of flows of functions and values as well as contaminants across county boundaries. Samples collected at the outlet (green circle) will contain signals from pesticides, functions and values originating outside of County A.

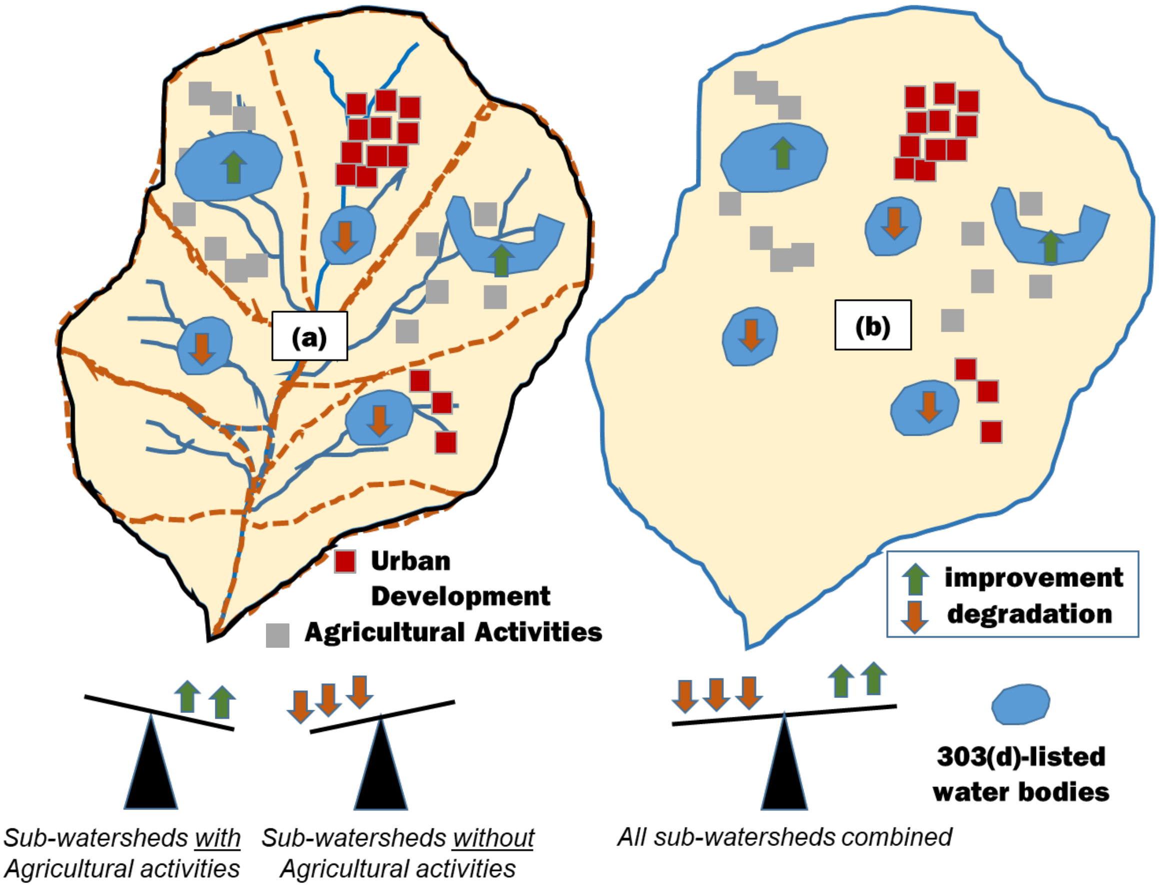


One approach to estimate the loading of various agriculturally-related contaminants is through modeling. Models, such as OpenNSPECT[45] utilize local and regional contextual information – e.g., slope steepness, soil type, land-use, and precipitation – to map pollutant loadings across watersheds. Models such as these can help estimate the expected contribution of various agricultural activities to specific functions and values. These types of estimates aid practitioners in assessing how land treatments or BMP's are likely to affect the "signal" detected by monitoring. An example of this approach was conducted in Whatcom Co. and applied to surface water in the Tenmile watershed[46]. Results of this effort facilitated the prioritization of land treatments and/ or BMP's within multiple sub-watersheds to achieve specific outreach and implementation goals pertaining to agricultural activities. The approach also helps practitioners disentangle the "signal" that is expected from agriculture from other influences beyond the boundary of VSP monitoring. A review and summary of related models and tools was published by American Farmland Trust in 2020[47].

Further consideration of monitoring boundaries is useful in the context of scale – that is, how to determine the scale boundaries most useful for your VSP work plan. Monitoring results can communicate varying patterns depending on the scale that they are assessed and reported. Consider that watersheds are multi-scale in nature. One larger basin may be composed of dozens of watersheds and hundreds of sub-watersheds. As such, smaller scale sub-watersheds can be aggregated into larger watersheds or basins. Depending on the underlying questions, it may be critical to ensure your monitoring approach is focused on the scale that best matches the goals and benchmarks of your county's watershed work plan. Aggregating data to larger scales may mask trends and relationships important for VSP (Fig. 27). Smaller scale data may improve the ability to isolate the signal from agricultural activities – which is desirable for informing VSP management, prioritization, and reporting. When working with "free" data provided to your watershed work group, it is essential that the information is evaluated to determine the most appropriate scale for use in your monitoring program.

For the most part, monitoring boundaries will vary by county based on the specific goals and objectives outlined in the county's work plan and/or adaptive management plan. In general it is useful to consider that samples and data are most useful for VSP when they are tightly linked to goals and benchmarks (i.e., "what is your question?"), and thus within the boundaries of a monitoring plan. While water-related functions and values have clear upstream/downstream flows due to physics and the behavior of hydrological systems, ecological-related functions and values often have different boundary constraints – such as distinct spatial areas of the landscape for monitoring, specific features of habitat necessary to quantify, or disentangling local drivers of conditions from regional, or global drivers of conditions.

Figure 27. Illustration of two approaches to monitoring: (*a*) separated by sub-watershed, and (*b*) aggregated to larger watershed. At smaller scales, such as the sub-watershed, it is easier to isolate the signal from agricultural activities for management, prioritization, and reporting. Blue polygons represent 303(d)-listed water bodies with directional arrows indicating the water bodies have improved or degraded over time. In panel (*a*), conditions are improving in the sub-watersheds containing agricultural activities but declining in the sub-watersheds without agricultural activities. In panel (*b*), it appears that overall conditions are declining, but without proper contextual information it is impossible to see that the declines are not driven by agricultural activities.


(a)
(b)
Urban Development
Agricultural Activities
improvement
degradation
303(d)-listed water bodies
Sub-watersheds with Agricultural activities
Sub-watersheds without Agricultural activities
All sub-watersheds combined

For example, if one of the objectives of the watershed work plan is to maintain or enhance pollinator function near agricultural areas, sampling for pollinators, indicators of pollinators, or pollinator habitat, will likely need to occur in areas near agricultural activities. For managed pollinators, e.g., the honey bee, the focal monitoring boundary may be tightly coupled to agricultural activities and the areas very near managed honey beehives. Such an approach may also assess the use of pesticides and/or the quality of pollen and nectar food resources in and around farms. If the focus is on wild pollinators, the monitoring boundary may need to be augmented to include different types of habitats near or far from the agricultural activities to understand how sampling near or within agricultural areas is affected by influences throughout the landscape in other areas. Within the domain of pollinator monitoring, the specifics of how, where, and what is sampled may differ depending on the underlying question (e.g., honey bees vs. wild pollinators). In this sense, it's useful to ensure the goals, benchmarks, and questions are well defined to aid in defining relevant boundaries of your monitoring program.

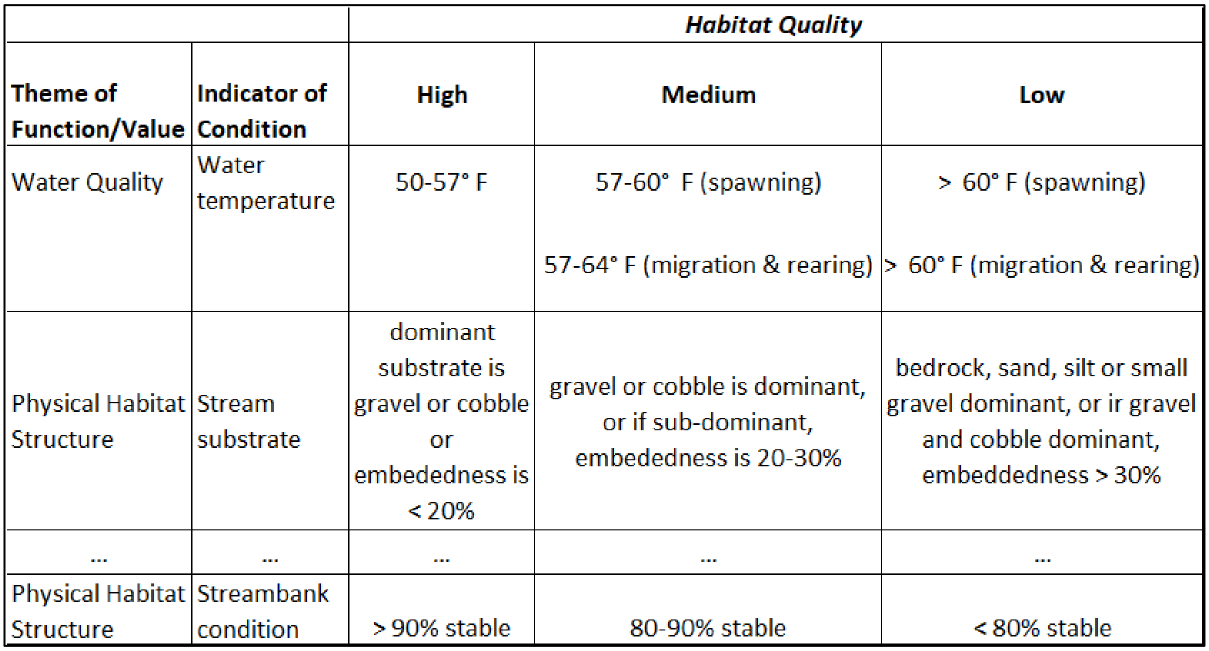
Boundary considerations are also key considerations for functions and values related to fish and wildlife. Anadromous fish spend time in oceans; ocean conditions are outside county-level control and are affected by a suite of factors unrelated to agricultural activities within a county – e.g., changes in ocean circulation patterns, temperature, and acidity. Thus, relying on benchmarks associated with anadromous fish counts may reflect the signal of ocean conditions rather than the signal of agricultural-related factors within a county. It may be useful to instead focus monitoring efforts and benchmarks on elements that are within the domain of county-level management activities, such as habitat quantity and quality.

Habitat Quantity, Quality, and Functions and Values

Habitat refers to the nexus of biotic and abiotic characteristic required for an organism to persist and reproduce (i.e., niche) – e.g., temperature, soil type, food resources, or nesting resources. Thus, the designation of habitat vs. non-habitat informs an additional monitoring boundary. *Habitat quantity* refers to the total amount (i.e., area or volume) of a specific habitat. For example, wetland-dependent waterfowl require wetlands for species persistence. Thus, the quantity of habitat for these birds may be assessed by counting the number and/or size (e.g., total acres) of wetlands within a watershed. Habitat quantity is often determined from combinations of field assessments and mapping approaches. For many species, habitat size is a key factor for persistence (i.e., sustainable populations through time) and the maintenance of biodiversity. Generally, larger habitats support higher levels of biodiversity[48] and offer a higher likelihood of persistence for species[49], though small habitat patches are still important for conservation[50]

Habitat quality refers to the quality of habitat characteristics for sustaining a species. Note that habitat quality is synonymous with habitat integrity and habitat condition (see *What are we monitoring? Relationship between functions & values, indicators, condition, quality, and integrity*). High quality habitats exhibit more favorable characteristics for species persistence than lower quality habitats. For example, a stream may be habitat for one or more species of fish, but for many fish species, warmer streams are a lower quality habitat compared to cooler streams (Tbl. 6). Wetlands may be habitat for a species of waterfowl, but some wetlands may provide more food or nesting resources than others because differences in the structure or composition of aquatic vegetation, adjacent upland vegetation, or hydrologic regimes. Habitat quality can be measured, ranked, or scored based on field assessments, models, samples, or combinations of related monitoring tools. Habitat for fish and wildlife is considered to be a *function* and/or *value* (see *Volume I: Critical area types and functions*). The function/value of habitat is underpinned by the composition and structure of habitat elements (e.g., water temperature, vegetation cover, soil depth, etc). Thus, a reduction in habitat quality (i.e. integrity or condition) will result in the degradation of *functions and values* related to habitat, and *vice versa*. Changes in habitat quantity and/or quality are often strong drivers of species population patterns (e.g., declines or increases) and are frequently related to activities occurring on the landscape, e.g., agricultural activities[51].

Table 6. Illustration of variable habitat qualities for salmon (adapted from the National Marine Fisheries Service[52]). Note that habitat quality is a function of multiple factors and can be assessed on a continuum.



		Habitat Quality		
Theme of Function/Value	**Indicator of Condition**	**High**	**Medium**	**Low**
Water Quality	Water temperature	50-57° F	57-60° F (spawning) 57-64° F (migration & rearing)	> 60° F (spawning) > 60° F (migration & rearing)
Physical Habitat Structure	Stream substrate	dominant substrate is gravel or cobble or embededness is < 20%	gravel or cobble is dominant, or if sub-dominant, embededness is 20-30%	bedrock, sand, silt or small gravel dominant, or ir gravel and cobble dominant, embeddedness > 30%
...	...	...	...	...
Physical Habitat Structure	Streambank condition	> 90% stable	80-90% stable	< 80% stable

In general, information on habitat quantity is typically easier to access and report on than information on habitat quality. Many practitioners are familiar with utilizing land-use/land-cover information, such as from DFW's High Resolution Change Detection

(HRCD)[53] program or the National Land Cover Database[54] from the USGS. In these and similar datasets a map pixel is often assigned a category, such as "woody wetland" or "grassland/herbaceous". Such designations, while often informative, lack information on habitat *quality* – which may be critical to know depending on the underlying questions informing your monitoring.

If, for example, your watershed work plan questions pertain to the persistence of a species with known habitat requirement, it may be desirable to determine trends in habitat quality and related habitat thresholds (i.e., similar to the habitat classifications in Tbl. 6). In such a scenario, habitat quality could be assessed and compared over time, similar to the depiction in Fig. 4.

Consider that it is possible for habitat *quantity* to remain unchanged (100 acres of wetlands in 2011 vs. 100 acres of wetlands in 2016) but for habitat *functions and values, condition,* or *integrity* to decline (e.g., 100 acres of *high functioning* wetlands in 2011 vs. 100 acres of *low functioning* wetlands in 2016). The need for information relating to habitat functions should be informed by the questions, goals, and benchmarks outlined in your watershed work plan and/or adaptive management plan.

Landscape and Habitat Connectivity

Habitat connectivity is another key factor related to fish and wildlife functions and values. Habitat connectivity refers to the degree that landscape elements impede or facilitate the movement of organisms or ecological processes. Consider that a highly fragmented landscape contains discrete patches of habitat (e.g., shrub-steppe, wetlands, streams, or native forest) within a "matrix" of non-habitat (e.g., roads, urban areas, or farms) (Fig. 28). What constitutes "habitat" depends on the species or community of interest to your watershed work group and/or the goals and benchmarks outlined in your watershed work plan. The terms "patch", "matrix", and "corridor" are used commonly by ecologists to refer to fragmented landscapes where the focal habitat(s) ("patch" or "patches") are relatively homogeneous ecological units (e.g., a patch of old growth forest or shrub-steppe) within a complex landscape. The "matrix" refers to the dominant landscape elements in the region in which the patches are situated (e.g., farmlands surrounding a forest patch). "Corridors" are linear patches that connect to other habitat patches within the matrix. More information about these concepts can be found in the linked resources section.[55]

A highly connected landscape may contain a large number of habitat patches that are structurally or functionally connected, such as via corridors. *Structural connectivity* refers to the degree that the habitat patches are physically near or adjacent to each other. For example, a contiguous wildlife corridor would have high structural connectivity, where a series of forest patches separated by urban areas would have low structural connectivity (Fig. 28). *Functional connectivity* refers to how connected the landscape is from the perspective of a species. Functional connectivity utilizes biological or behavioral attributes of a species to estimate how the landscape impedes or facilitates animal movement for migration, dispersal, or daily travel. Connectivity varies by species due to differences in dispersal and migration distances or daily travel. Consider that the dispersal or migration distance associated with a migratory bird (10 km – 1,000 km) is likely much larger than that of a frog (10 m – 1,000m). As such, functional connectivity levels may be different for different species within the same landscape. Reviews of habitat connectivity and associated metrics provide more detail on connectivity considerations as well as tools and software to estimate landscape connectivity (Calabrese & Fagan 2004)[56,57].

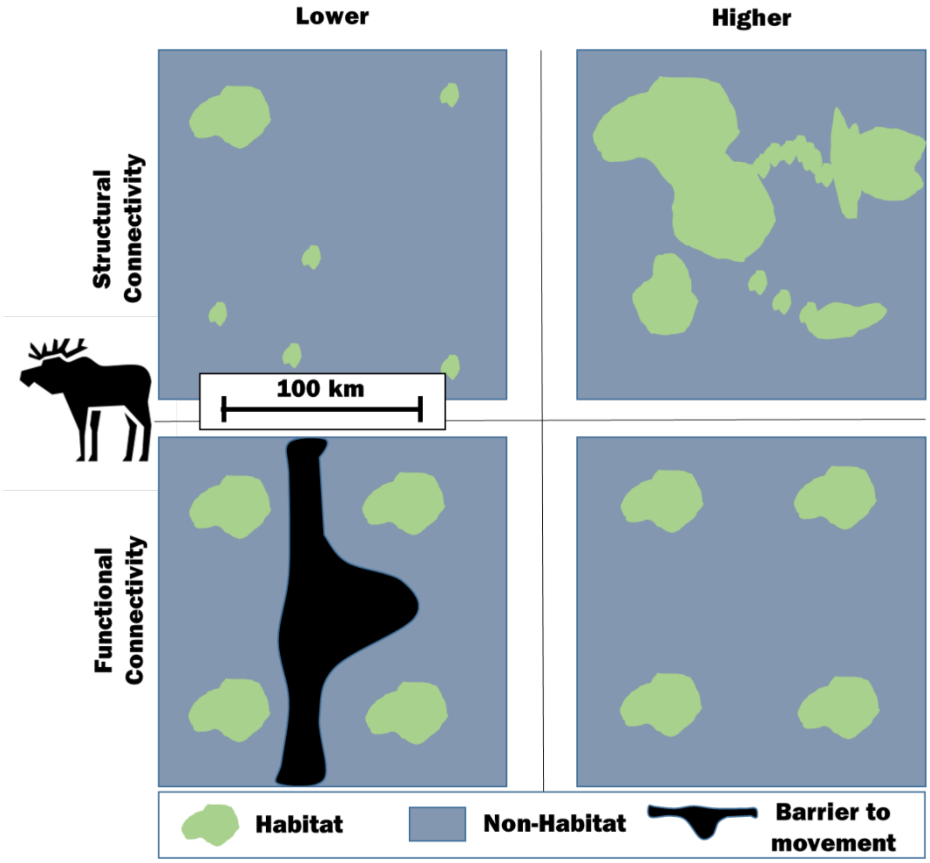
Within a watershed or county, there may be multiple landscape elements that impede or facilitate animal movement – including natural and human-created landscape features. In stream systems, barriers to fish connectivity are non-fish-friendly culverts (i.e., culverts that are narrow, have a high velocity of flow, and/or a waterfall out of the culvert) that prevent fish passage. Alternatively, fish-friendly culverts may facilitate fish movement. Depending on where culverts are located, they may have disproportionate effects on fish connectivity. Barriers in headwater streams may have less of an impact on connectivity than barriers in main stem streams simply due to their relative locations within the watershed.

It's important to remember that landscape connectivity is context dependent. Habitat for one species may be inhospitable to another species, and vice versa. Woodland-dependent species may traverse non-woodland areas in order to migrate between forest patches. Some woodland-dependent species avoid open areas, such as grasslands or farms, and thus in this context the distribution of forest patches, grasslands, and farms in a landscape can affect how connected the landscape is from the standpoint of the focal species (Fig. 28).

Agricultural practices may facilitate or impede animal movement. For example, hedgerows, intact riparian areas, and remnant habitats within an agricultural landscape may facilitate connectivity for some species, such as butterflies, bats, and birds[58,59]. Alternatively, some species, such as sage grouse, bighorn sheep, and pygmy rabbits may avoid movements in and around farms[60], in which case the extent and distribution of farms relative to native habitat may impede landscape connectivity.

If your watershed work plan activities include a focus on landscape connectivity, one approach that may be useful is to develop one or more indicators or metrics of landscape connectivity that can be compared over time (see *Detecting changes over time: what direction are things headed?*). A suite of landscape connectivity metrics and approaches can be found in Calabrese & Fagan 2004 and references *56*, *57*, and *58* in the *Linked References and Resources* section. If your work plan does not include landscape connectivity in this context, coordinate with your local WDFW habitat biologist to determine whether updating your plan might be needed.

Figure 28. Depiction of structural and functional connectivity for a hypothetical moose species. Distinct habitat patches are separated by an "ocean" of non-habitat. Landscape and habitat connectivity characterize landscape feature to determine how easily wildlife can traverse the landscape for ecological needs. Structural connectivity increases when habitat patches are closer together or connected. Functional connectivity increases when barriers to fish and wildlife movement are reduced.



Sample timing and frequency

Timing is a critical aspect of monitoring. From the standpoint of data collection, determining when samples or monitoring information is to be gathered is key to understanding the subject of interest – e.g., seasonality (summer vs. winter), pre- or post-precipitation event, or when specific plants are in bloom. Timing also plays a role in establishing trends in various phenomena given that there are often lag-times associated with ecosystem functions and values (Fig. 29).

Remember to re-ask: “what is my question?”. Consider how time-of-year, time-of-day, or other time-related considerations may influence the characteristics that are of most importance to the underlying question(s). Some examples are provided below.

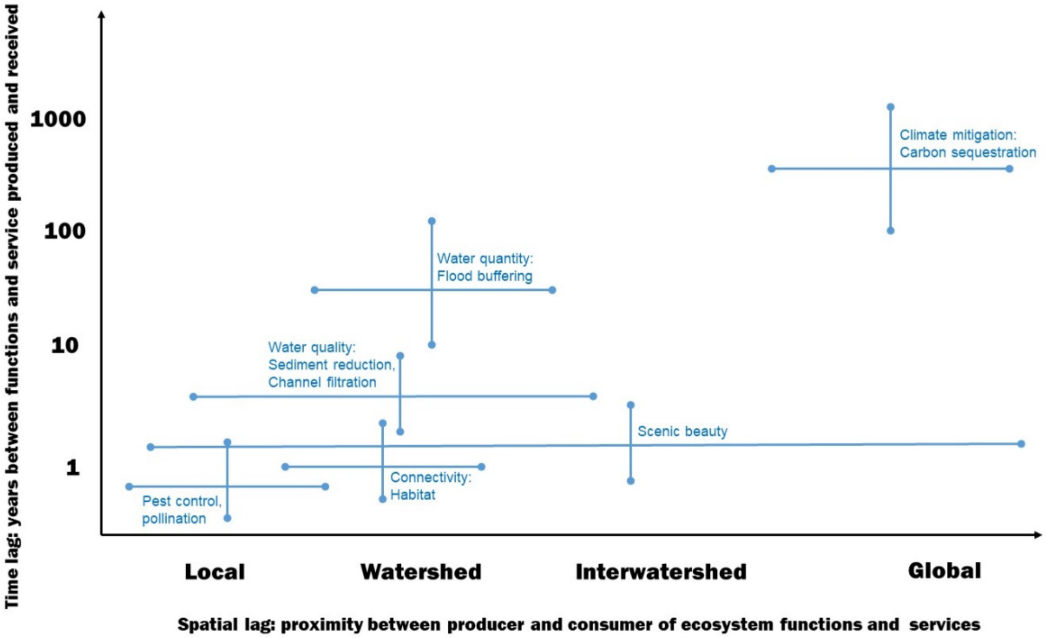
Bacterial loading in surface water can respond to rainfall events up to 3 weeks prior to sampling and respond differently depending on the volume of rainfall (Leight and Hood 2018). If watershed work plan goals and benchmarks include monitoring surface water bacteria, the monitoring plan should explicitly outline how sampling accounts for rainfall given the strong influence of rainfall on bacteria.

Similarly, many streams in Washington experience critical summer periods, where stream flows and water volumes are typically lower than other times of the year. Low stream flows are associated with higher stream temperatures, lower dissolved oxygen, and higher concentrations of contaminants[61], potentially affecting the quality, condition, or integrity underlying the functions and values of interest to watershed work groups (Tbl. 6). However, surface water samples during the summer critical period may exhibit less pronounced changes in temperature and dissolved oxygen if they are collected from locations influenced by groundwater, which tends to be much cooler than surface water. Conversely, a question related to high summer stream temperature would not necessarily need to sample winter storm runoff. Moreover, the characteristics of hillslopes and watersheds draining into the surface waters of interest often affect stream flows. Understanding the context of the watershed, its current and historical trends and conditions, critical environmental constraints and the nature of the agricultural activities present, will aide in developing sensible timeframes for monitoring.

Further, it’s useful to consider that it takes time to see a change in the direction of a natural resource-related signal in response to land treatments or BMPs. It may take up to 10 or 50 years to “see” changes in some signals of condition, while other signals may respond within a few years (Fig. 29) (Fremier *et al.* 2013). The time lag in response to land treatments may be incorporated into monitoring activities and the interpretation of monitoring results. For example, it may take years to detect the signal of decreased agricultural-related erosion in surface water samples due to hydrologic sediment deposition and flushing times in streams (Brooks *et al.* 2010). Arming the watershed work group with reasonable expectations with respect to the timing of perceived responses of critical area conditions will improve reporting of monitoring results.

The anticipated effects of land treatments on critical area conditions may be estimated using previous projects or modeling approaches. Estimates and information from these types of approaches should help inform your watershed work group of reasonable expectations for, and interpretation of, monitoring results. While the time scale for detecting shifts in the signal of interest may be rather long for some phenomenon, it’s crucial to continue monitoring on a regular schedule, over time, to ensure that adequate data is collected and available for making inferences about the effectiveness of VSP activities in your county. Lack of data greatly limits the ability of work groups to adapt to changing conditions or to better align implementation of conservation practices with shifting natural resource concerns.

Figure 29. Illustration of lags between the production and consumption of ecosystem functions and services across spatial scales and timeframes (adapted from Fremier *et al.* 2013). In the context of monitoring, it is critical to understand how time lags may influence the types of information being collected or sampled. For example, it may take years to detect the signal of decreased agricultural-related erosion in surface water samples due to hydrologic sediment deposition and flushing times in streams (Brooks *et al.* 2010).



Operationalizing a Monitoring Program

Monitoring is both a science and an art, in that it requires a foundation of techniques and approaches that is attended-to and iteratively improved with a "vision" for the end result. Some subjectivity is unavoidable within the domain of monitoring given the myriad of possible constraints – e.g., access to sampling locations and data, availability of geographic contextual information, changing socio-environmental conditions, and staffing considerations. Monitoring practitioners within VSP are encouraged to reflect on the content laid out within this guide in the context of their watershed work plans in order to refine monitoring approaches. It is hoped that diligent framing of VSP monitoring within the outline of key monitoring principles, such as described in Fig. 2, will lead to the development of robust monitoring plans that employ methods and approaches well-suited to the questions, goals, and benchmarks of most importance to their county's watershed work plan. If the foundational principals of monitoring outlined in this guide are understood and followed, the monitoring plan should be sufficient to: (i) conclusively determine answer(s) to overarching questions, (ii) communicate the level of confidence in the answer(s), (iii) inform VSP management activities, e.g., prioritizing land treatments and/or BMPs, to enhance capacity for adaptive management, (iv) identify clearly the boundaries of what is, and what isn't, within the scope of the monitoring plan (i.e., set appropriate expectations for monitoring outcomes), and (v) provide a framework for collaborative relationships with other staff and organizations via a common language for monitoring at the watershed scale.

The programmatic guardrails for monitoring within VSP are established from the VSP statute and addressed in the *Overview of VSP* section. Key aspects include a focus on monitoring at the watershed scale, the role of watershed work groups as decision-makers regarding planning, prioritization, and monitoring, and the use of adaptively managed watershed work plans to communicate goals and benchmarks pertaining to the protection and enhancement of ecosystem functions and values.

Balancing the need for robust, actionable monitoring for VSP reporting with privacy concerns of stakeholders can be challenging. However, because monitoring results are required at the watershed scale – not the parcel or farm scale – there are approaches to monitoring that can achieve this balance with thoughtful consideration. It is hoped that material from this guide can inform these considerations, such as trade-offs between monitoring at sub-watershed vs. watershed scales (Fig. 27) and the locations used for sampling relative to the signal of agricultural activities and the questions of most importance to the watershed work group (Figs. 8, 9, and 25). Moreover, the choice of what to sample may also present trade-offs for stakeholder privacy (Tbl. 1).

Monitoring within VSP allows for the use of information from a broad range of sources, including (but not limited to) conservation districts, county departments, state agencies, universities, tribes, federal agencies, community groups, and non-governmental organizations. Often information made available from these sources is not intrinsically aligned with VSP goals. When working with "free" data provided to your watershed work group, it is essential that the information is evaluated to determine whether it is informative in light of the goals, benchmarks, and questions outlined in your watershed work plan.

Evaluation of such information should consider:

(i) Whether the information informs a pertinent question. Are the goals, objectives, and questions clearly articulated in the watershed work plan? Does the ("free") data provided inform the stated goals, objectives, and questions outlined in the work plan?

(ii) The scale of the information provided – i.e., can we answer questions at the watershed scale with the information? Does the information raise privacy concerns among stakeholders?, Can we scale-up or scale-down the information to better match our needs (Fig. 27)?

(iii) The accuracy and precision of the information in light of the do-something-level identified by the watershed work plan – i.e., is the accuracy/precision level sufficient to answer our questions with certainty?

(iv) The relationship between the information provided and the functions and values of interest the watershed work group – i.e., does the information relate directly or indirectly to functions and values of concern?

(v) The location and timing of the information provided – i.e., was it collected in locations that are meaningful (e.g., upstream vs. downstream, in a habitat of concern)? Was it collected in a timeframe that is meaningful (e.g., 20 years ago vs. 5 years ago, after a rainfall event vs. before a rainfall event)? Is it detecting the potential "signal" of agricultural activities?

(vi) How the information was collected or determined – i.e., the monitoring approach used (total suspended solids determined in a lab vs. settleable solids from an Imhoff cone; Tbl. 1). Does the method of collection match the needs for monitoring outlined in the watershed work plan?

(vii) The completeness of the information provided – i.e., is there a sufficient number of samples collected to conclusively answer the question(s)? If not, can we use what data exists to determine how much more monitoring is needed to answer the question?

Ultimately, decisions about monitoring approaches should be informed by balancing the questions of most importance to the watershed work group (outlined in the watershed work plan), the scales most relevant for the work group questions (keeping in mind the watershed-scale focus of VSP), the practical constraints of monitoring – such as cost, staff time, data availability, and stakeholder privacy, and the ability for the monitoring data to inform VSP activities – such as planning, prioritization, adaptive management, and reporting. In some cases information provided to the watershed work may not meet all of the needs of the watershed work group – but may still provide useful context and/or may be used to develop additional monitoring activities that are more targeted to VSP.

Foundational principals of monitoring are addressed throughout the guide, beginning with setting the stage – focusing on clear goals, objectives, and questions; the themes of interest – such as effectiveness (e.g., for functions and values), implementation, or participation; and the agricultural and ecosystem context that “drives the system”.

Once the stage is set, it is appropriate to consider the hypotheses to be addressed and the role that statistics play in determining whether the monitoring results are “better than random” – an essential aspect of reporting. This is done by highlighting three key aspects of monitoring: signal, noise, and errors – and assessing monitoring options in light of a “do-something-level”. Combining these factors allows for practitioners to estimate how much data is needed and how frequently it needs to be collected to conclusively answer questions (see *Determining how many samples you need to answer your question: Power analysis*). The essential elements of how to analyze this information to compare current conditions with the VSP baseline (July 2011) are outlined in *Detecting changes over time: what direction are things headed?*

This guide distills essential monitoring elements as concisely as possible and, as such, does not convey the depth and detail required for all monitoring needs. However, references contained within the guide are meant to direct practitioners to additional resources where necessary. Excellent resources containing more in-depth information on a variety of topics addressed within the guide, such as more detailed treatment of statistical considerations, sampling considerations, decision trees, and work flows, can be found in EPA's publications "Monitoring and Evaluating Nonpoint Source Watershed Projects" [62] and "Monitoring Guidelines to Evaluate Effects of Forestry Activities on Streams in the Pacific Northwest and Alaska" [63]. Additionally, NRCS's Conservation Effects Assessment Project (CEAP) program has published several summaries of insights gained from watershed-scale monitoring of agricultural activities[64,65]. Further resources can be found within the *Linked References and Resources* section and are cited throughout the document.

Finally, developing a monitoring *plan* from the various elements of VSP monitoring outlined in this guide will facilitate the collection, analysis, and reporting of information for VSP. A forthcoming section (*Volume III: Developing a monitoring plan*) will provide a basic framework for developing such a plan. Examples of similar approaches can be found in resources from ECY[66], EPA[67], NRCS[68], and USGS[69]. A clear monitoring plan will help with day-to-day activities by essentially "doing the work up front" – in that it is easier to delegate, communicate, and operate when the *why*, *who, what, where,* and *when* of monitoring is outlined and scheduled in advance. If the themes of this guide are followed, and a monitoring plan from these themes is developed, practitioners should have confidence that monitoring activities will result in actionable information that is appropriate for reporting and management within VSP. Moreover, a monitoring plan also sets the expectations of what is possible to ascertain through the monitoring activities – potentially alleviating conflicts that may arise when VSP stakeholders or practitioners have misaligned expectations for what claims monitoring information can, and can't, support within VSP.

Definition of Terms and Acronyms

Acronyms

BMP: Best Management Practice – “include(s) soil and water conservation practices, other management techniques, and social actions that are developed for a particular region as effective and practical tools for environmental protection”[70]

DFW: Washington State Department of Fish and Wildlife

DNR: Washington State Department of Natural Resources

ECY: Washington State Department of Ecology

EPA: US Environmental Protection Agency

FWHCA: Fish and Wildlife Habitat Conservation Area

GMA: Growth Management Act

ISP: Individual Stewardship Plan

NTU: Nephelometric Turbidity Units. Unit of measure used for quantifying turbidity based on how light scatters in relation to particles in liquid.

SCC: Washington State Conservation Commission

TMDL: Total Maximum Daily Load. “A TMDL is the calculation of the maximum amount of a pollutant allowed to enter a waterbody so that the waterbody will meet and continue to meet water quality standards for that particular pollutant. A TMDL determines a pollutant reduction target and allocates load reductions necessary to the source(s) of the pollutant”. (**https://www.epa.gov/tmdl/overview-total-maximum-daily-loads-tmdls**; accessed 11/22/2021)

TSP: Technical Service Provider

USGS: U.S. Geological Survey

VSP: Voluntary Stewardship Program (SCC).

WRIA: Water Resource Inventory Area

WSDA: Washington State Department of Agriculture

Definitions

Box and Whiskers plot: a chart developed from data that displays key elements of the information collected, including the minimum, median, and maximum values obtained. These plots are useful for exploratory analysis and confirming underlying data assumptions related to statistical tests used in monitoring[71].

Condition (environmental): An environmental condition is synonymous with environmental quality or integrity. Condition refers to the condition of a feature (e.g., a Critical Area) relative to a similar feature in a reference (pristine) condition. For example, a diked wetland adjacent to roads would have lower integrity, quality, or condition compared to a wetland located within a pristine nature preserve with no dikes or roads.

Corridor: a linear “patch” that connects other discrete habitat patches within a “matrix” of non-habitat. Corridors confer connectivity for organisms via enhanced structural connectivity of landscapes.

Explanatory Variable: one or more variables (number of conservation practices, elevation, temperature, time [years, months]) that explain how a response variable changes. An explanatory variable is what you are using to explain how the response variable changes. Ex. Precipitation (explanatory variable) directly relates to our soil wetness measurements (response variable). Note: Explanatory variables are also referred to as “independent variables”.

Functions: functions manifest from interactions between the structure and composition of living (biotic) and non-living (abiotic) elements of an ecosystem. These interactions often occur at multiple scales from the very small (genes or molecules) to the very large (landscapes). Examples include: flood storage, erosion prevention, and water quality improvement.

Indicator: Indicators are used to measure composition or structure and/or the quality, condition, and/or integrity of some environmental feature, e.g., a Critical Area. There are a wide range of indicators that have been developed for specific purposes. Examples include *E. coli* (fecal contamination), macro-invertebrates (biological condition of surface waters), or indices of light reflectance from satellites, such as the Normalized Difference Vegetation Index (NDVI; vegetation growth or condition).

Imhoff Cone: a conical container used to measure the volume of settle able solids in a standardized manner.

Integrity: Environmental integrity is synonymous with environmental quality or condition. Integrity refers to the condition of a feature (e.g., a Critical Area) relative to a similar feature in a reference (pristine) condition. For example, a diked wetland adjacent to roads would have lower integrity, quality, or condition compared to a wetland located within a pristine nature preserve with no dikes or roads.

Matrix: dominant landscape element that contrasts with a discrete “patch” of focal habitat. The “matrix” is the ocean of non-focal habitat that surrounds focal patches of habitat within a landscape, such as agricultural fields surrounding a patch of native forest.

Normal Distribution: a bell-shaped curve that arises when the distribution of data values (e.g., collected through sampling) is plotted on the x-axis in relation to the frequency that a value was encountered on the y-axis. Also referred to as a Gaussian distribution. See Appendix-1 for more information.

Patch (habitat): a discrete ecological unit that is relatively homogeneous and contrasts with the features of the landscape as a whole, such as a patch of native forest surrounded by grasslands.

Quadrat: a square shaped boundary used to standardize on-the-ground samples or observations, such as plant diversity or soil characteristics.

Quality: Environmental quality is synonymous with environmental condition or integrity. Quality refers to the condition of a feature (e.g., a Critical Area) relative to a similar feature in a reference (pristine) condition. For example, a diked wetland adjacent to roads would have lower integrity, quality, or condition compared to a wetland located within a pristine nature preserve with no dikes or roads.

Response Variable: the focal subject of interest (e.g., ecosystem functions & values, participation, etc.) that “responds” to one or more explanatory variables. A response variable is what you are solving for in a statistical equation. Ex. Surface water nutrients (response variable) increase with fertilizer use (explanatory variable). Note: response variables are also referred to as “dependent variables”.

Turbidity: a measure of liquid clarity. Turbidity is measure by measuring how light penetrates or reflects liquids in response to material within the liquid. Within VSP turbidity may be used to measure particulate matter in surface waters.

Values: Values represent the benefits of functions to society, culture, and economies. Often, values are considered synonymous with “ecosystem services”.

References

Brooks, E.S., Boll, J., Snyder, A.J., Ostrowski, K.M., Kane, S.L., Wulfhorst, J.D., Van Tassell, L.W. and R. Mahler. 2010. Journal of Soil and Water Conservation. 65(6): 331-341. DOI: 10.2489/jswc.65.6.331

Calabrese, J.M. and W.F. Fagan. 2004. A comparison-shopper's guide to connectivity metrics. Frontiers in Ecology and the Environment. 2(10): 529-536. http://science.umd.edu/biology/faganlab/pdf/CalabreseFagan2004b.pdf

Fremier, A.K., DeClerck, F.A.J., Bosque-Pérez, N.A., Estrada Carmona, N., Hill, R., Joyal, T., Keesecker, L., Klos, P.Z., Martínez-Salinas, A., Niemeyer, R., Sanfiorenzo, A., Welsh, K. and J.D. Wulfhorst. Understanding Spatiotemporal Lags in Ecosystem Services to Improve Incentives. BioScience. 63(6):472–482. https://doi.org/10.1525/bio.2013.63.6.9

Halsey, L. G. 2019. The reign of the p-value is over: what alternative analyses could we employ to fill the power vacuum? Biology Letters.15: 20190174. 20190174 http://doi.org/10.1098/rsbl.2019.0174

Leight, A.K. and R.R. Hood. 2018. Precipitation thresholds for fecal bacterial indicators in the Chesapeake Bay. Water Research. 139: 252-262. https://doi.org/10.1016/j.watres.2018.04.004

Lichstein, J.W., Simons, T.R., Shriner, S.A. and Franzreb, K.E. 2002. Spatial autocorrelation and autoregressive models in ecology. Ecological Monographs. 72: 445-463. https://doi.org/10.1890/0012-9615(2002)072[0445:SAAAMI]2.0.CO;2

Murtaugh, P. A. 2014. In defense of P values. Ecology. 95:611–617. DOI:10.1890/13-0590.1

Zuur, A., Ieno, E. N., Walker, N., Saveliev, A. A., & G. M. Smith 2009. Mixed effects models and extensions in ecology with R. Springer Science & Business Media.

Zuur, A.F., Ieno, E.N. and C.S. Elphick. 2010. A protocol for data exploration to avoid common statistical problems. Methods in Ecology and Evolution, 1: 3-14. https://doi.org/10.1111/j.2041-210X.2009.00001.x

Tobler W. 1970. "A computer movie simulating urban growth in the Detroit region". Economic Geography, 46(Supplement): 234–240.

List of Figures

Figure 16. Example of statistically significant trend lines for a time series plot.

Figure 17. Example of trend lines that *are not* statistically significant for a time series plot.

Figure 18. Conceptual illustration of a time series plot divided into baseline and comparison categories.

Figure 19. Conceptual illustration of a time series plot and associated histograms for baseline and comparison categories.

Figure 20. Conceptual illustration of a time series plot and associated histograms for baseline and comparison categories with high levels of “noise” (variability).

Figure 21. Conceptual illustration of a statistically significant time series plot and associated histograms for baseline and comparison categories with low levels of “noise”.

Figure 22. Example of one-tailed tests applied to dissolved oxygen over time relative to baseline conditions. *(a)* Comparison central tendency is greater than the baseline central tendency. *(b)* Comparison central tendency is less than the baseline central tendency.

Figure 23. Example of a two-tailed test applied to dissolved oxygen over time relative to baseline conditions.

Figure 24. Graphical results of a worked example for power analysis. *(a)* Plot of mean dissolved oxygen concentrations (colored bars) with associated “noise” (standard deviation lines) for one station, for two time periods: baseline (2006-2011) and comparison (2011-2016). *(b)* Histogram plot of dissolved oxygen exhibiting an approximately normal distribution. *(c)* Time series plot showing the trend in dissolved oxygen concentration over time for one surface water station with 50% and 95% confidence levels indicated. *(d)* Results of a power analysis with the number of samples needed to see a specific change from baseline dissolved oxygen. *(e)* Time series plot showing the trend in dissolved oxygen concentration over time for a subset of only 20 samples from the original data.

Figure 25. Conceptual illustration of soil erosion occurring from agricultural activities and urban development.

Figure 26. Illustration of flows of functions and values as well as contaminants across county boundaries.

Figure 27. Illustration of two approaches to monitoring: (*a*) separated by sub-watershed, and (*b*) aggregated to larger watershed.

Figure 28. Depiction of structural and functional connectivity for a hypothetical ruminant animal.

Figure 29. Illustration of lags between the production and consumption of ecosystem functions and values across spatial scales and timeframes.

Figure A-1. A histogram of temperature samples from a reasonable accurate and precise thermometer.

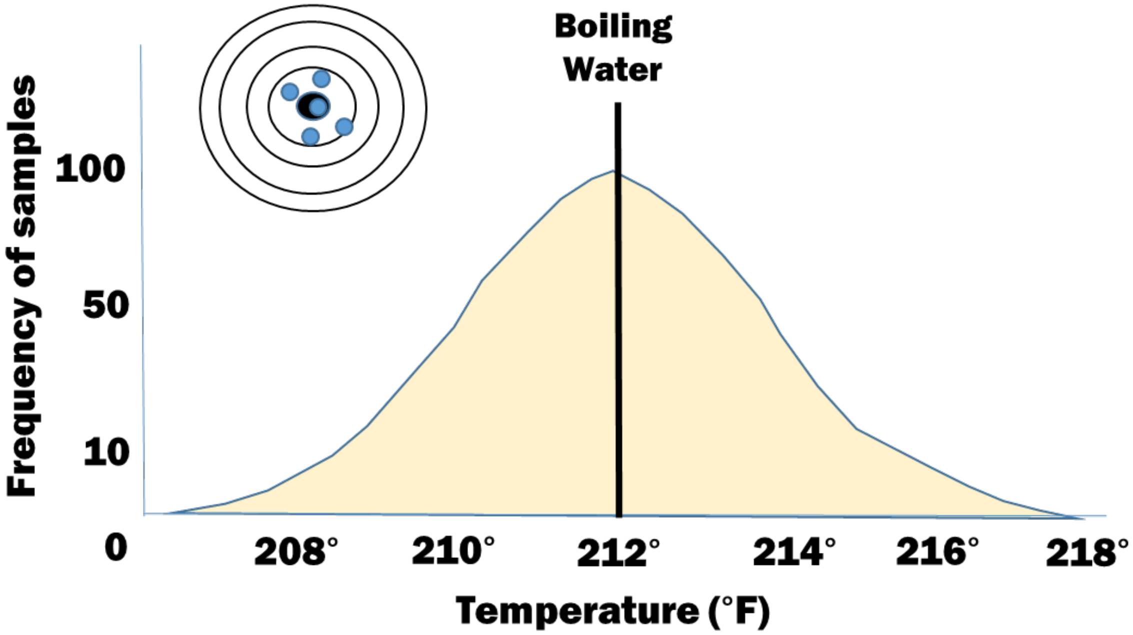
Figure A-2. A histogram of temperature samples from an inaccurate and imprecise thermometer.

List of Tables

Appendices

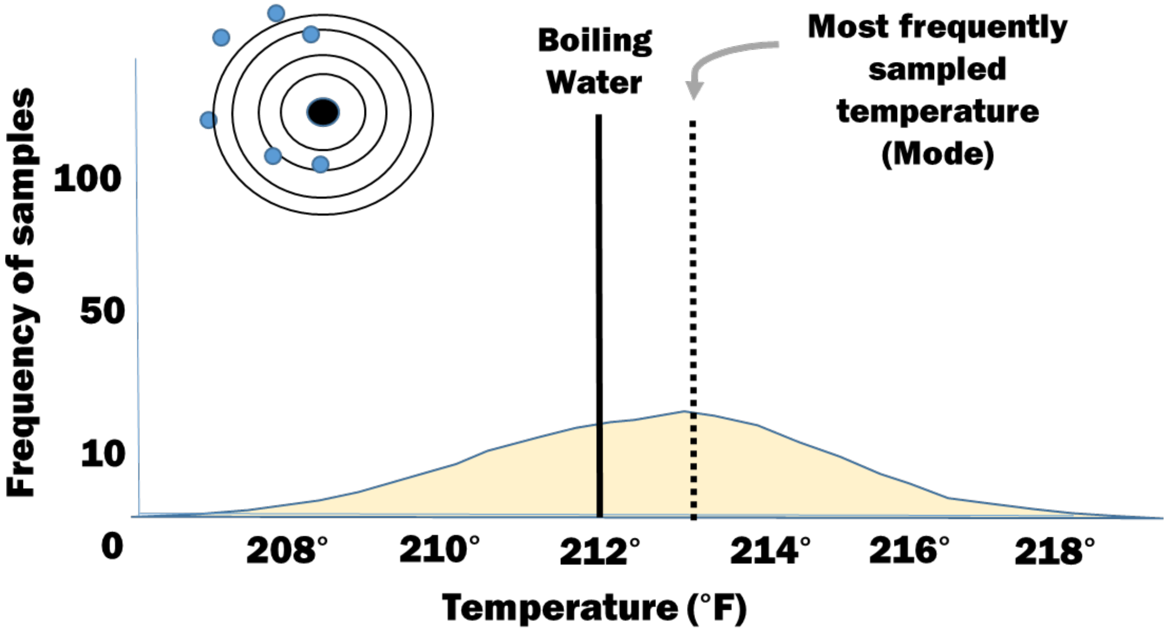
1-Histograms

Figure A-1. A histogram of temperature samples from a reasonable accurate and precise thermometer. The noise of the data is relatively modest, clustered around the true (known) temperature of boiling water. The shape of the histogram indicates the data are normally distributed – indicated by the "bell-shape".



Note that in Fig. A-1, the target shows points clustered around the center, with little variation between data points. This is reflected in the histogram as showing a clear central tendency right at 212°F with a bell-shaped curve on either side. This illustrates temperature measurements from our accurate, precise, thermometer.

Figure A-2. A histogram of temperature samples from an inaccurate and imprecise thermometer. The noise of the data is relatively large, not clustered around the true (known) temperature of boiling water, but rather centered closer to 213°F. The shape of the histogram indicates the data are normally distributed – indicated by the "bell-shape". Wider spread, indicating more noise in the data, makes conclusive testing for differences more difficult.



If we use a thermometer that is less accurate and precise, as indicated by the points around the target in Fig. A-2, we will have more noise in our data and a less clear idea of our central tendency. This is indicated in the histogram of Fig. A-2 where the peak of the histogram is compressed relative the values around it. In general, data such as Fig. A-1 will be easier to utilize for making inferences compared to Fig. A-2, in part because the signal is clearer and the data variability is lower.

2-Table of Statistical Approaches

Parametric Approach	**Non-Parametric Equivalent**	**Purpose of Test**	**Example**
Paired t-test	Wilcoxon signed-rank test	Compare signals (i.e., means/ranks) of the same response (i.e., nutrients, no. of fish, temperature) at the same location/category at two different times to see if they are the same or different.	Compare average stream temperature at the same location before and after riparian restoration.
Two-sample t-test (a.k.a. Student's t-test)	Wilcoxon rank-sum test	Compare signals (i.e., means/ranks) of the same response (i.e., nutrients, no. of fish, temperature) from two different (independent) locations/categories to see if they are the same of different.	Compare average stream temperature between samples of two different streams.
Analysis of Variance (ANOVA)	Kruskal-Wallis test	Compare signals (i.e., means/ranks) between three or more locations/categories to see if they are the same or different (i.e., similar to the two-sample t-test but for situations with more than two samples).	Compare average stream temperature between samples of five different streams.
Regression	Spearman's rank correlation	Determine if the signal of interest (i.e., nutrients, no. of fish, temperature) is associated with another variable of interest (i.e., stream flow, steepness, time)	Explore an association between wetland biodiversity and turbidity – e.g., does wetland biodiversity decrease when turbidity increases?

Linked References and Resources

[1] What is a watershed? National Oceanic and Atmospheric Administration https://oceanservice.noaa.gov/facts/watershed.html

[2] Watersheds 101 (Utah State University) https://extension.usu.edu/waterquality/files-ou/Posters-and-Presentations/Watersheds_101_updated_6-5-2017.pdf

[3] Washington Department of Ecology: TMDL Water Quality Improvement Projects (organized by county) https://ecology.wa.gov/Water-Shorelines/Water-quality/Water-improvement/Total-Maximum-Daily-Load-process/Directory-of-improvement-projects

[4] USGS StreamStats web application https://streamstats.usgs.gov/ss/

[5] Upper Columbia Ecosystem Diagnosis and Treatment (EDT) https://www.ucsrb.org/edt/

[6] State of Our Watersheds (2020). Northwest Indian Fisheries Commission https://nwifc.org/publications/state-of-our-watersheds/

[7] Washington State Agland Database https://nrsig.org/projects/washington-state-agland-database

[8] Cascades to Coast Landscape Collaborative. 2021. Coastal Northwest Landscape Conservation Mapper. https://www.ctoclc.org/conservationresources

[9] Agricultural Conservation Planning Framework. USDA and Iowa State University. https://acpf4watersheds.org/

[10] Tenmile watershed assessment, Whatcom Co. (NRCS NWQI) https://www.whatcomcd.org/_files/ugd/74f091_f6499b3712af4cc38fda7c546d210163.pdf

[11] Tracking Conservation Tools (EPA). https://www.epa.gov/sites/default/files/2020-11/documents/tracking_conservation_tools_508.pdf?VersionId=l0QXtpMMsy83z2Z7Pu4c4v3.EHFqxd7_

[12] Michael A. Schroeder, R. C. Crawford, F. J. Rocchio, D. J. Pierce, and M. Vander Haegen 2011. Ecological integrity assessments: monitoring and evaluation of wildlife areas in Washington. Washington Department of Fish and Wildlife, Olympia, Washington. https://wdfw.wa.gov/publications/01314

[13] Washington State Department of Natural Resources – Ecological Integrity Assessment https://www.dnr.wa.gov/NHP-EIA

[14] Washington State Department of Ecology. Wetlands Rating Systems. https://ecology.wa.gov/Water-Shorelines/Wetlands/Tools-resources/Rating-systems

[15] Michael A. Schroeder, R. C. Crawford, F. J. Rocchio, D. J. Pierce, and M. Vander Haegen 2011. Ecological integrity assessments: monitoring and evaluation of wildlife areas in Washington. Washington Department of Fish and Wildlife, Olympia, Washington. https://wdfw.wa.gov/sites/default/files/publications/01314/wdfw01314.pdf

[16] Millennium Ecosystem Assessment. 2005. Ecosystems and Human Well-Being: Synthesis. Island Press, Washington, DC.
https://www.millenniumassessment.org/documents/document.356.aspx.pdf

[17] US Environmental Protection Agency. Indicators: Benthic Macro-Invertebrates.
https://www.epa.gov/national-aquatic-resource-surveys/indicators-benthic-macroinvertebrates

[18] US Geological Survey. NDVI, the Foundation for Remote Sensing Technology
https://www.usgs.gov/special-topics/remote-sensing-phenology/science/ndvi-foundation-remote-sensing-phenology

[19] Is the water dirty? Well, it depends on the bacteria standard.
https://rangelands.ucdavis.edu/is-the-water-dirty-well-it-depends-upon-the-bacteria-standard/

[20] Salmon Habitat Limiting Factors in Washington State
https://rco.wa.gov/wp-content/uploads/2019/10/GSRO-LimitingFactorReport.pdf

[21] Making Ecological Monitoring Successful
https://openresearch-repository.anu.edu.au/bitstream/1885/130886/1/Making%20ecological%20monitoring%20successful_web.pdf

[22] Monitoring Populations and their Habitats: A Practitioner's Guide. Chapter II. Data Analysis in Monitoring
https://open.oregonstate.education/monitoring/chapter/data-analysis-in-monitoring/

[23] Monitoring Guidance for Determining the Effectiveness of Nonpoint Source Controls (Ch. 4)
https://www.epa.gov/sites/default/files/2015-10/documents/monitoring_chap4_1997.pdf

[24] A biologist's guide to statistical thinking and analysis*
David S. Fay and Ken Gerow.
http://www.wormbook.org/chapters/www_statisticalanalysis/statisticalanalysis.html

[25] P values, hypothesis testing, and model selection: it's deja vu all over again.
https://harvardforest1.fas.harvard.edu/sites/harvardforest.fas.harvard.edu/files/publications/pdfs/Ellison_Ecology_2014.pdf

[26] Monitoring Guidelines to Evaluate Effects of Forestry Activities on Streams in the Pacific Northwest and Alaska
https://digital.lib.washington.edu/researchworks/bitstream/handle/1773/17068/Monitoring%20Guidelines%20to%20Evaluate%20Effects%20of%20Forestry%20Activities%20on%20Streams.pdf?sequence=1&isAllowed=y

[27] Handbook of Biological Statistics
http://www.biostathandbook.com/

[28] SSN: An R Package for Spatial Statistical Modeling on Stream Networks

https://www.jstatsoft.org/article/view/v056i03

[29] Applied Time Series Analysis for Fisheries and Environmental Sciences https://nwfsc-timeseries.github.io/atsa-labs/

[30] Autocorrelation and Time Series Methods https://online.stat.psu.edu/stat462/node/188/

[31] Quantifying temporal change in biodiversity: challenges and opportunities https://royalsocietypublishing.org/doi/pdf/10.1098/rspb.2012.1931

[32] Parametric and Nonparametric: Demystifying the Terms https://www.mayo.edu/research/documents/parametric-and-nonparametric-demystifying-the-terms/doc-20408960

[33] NIST/SEMATECH e-Handbook of Statistical Methods. Retrieved 2021. http://www.itl.nist.gov/div898/handbook/

[34] Hypotheses for paired and two-sample t-test means. https://www.khanacademy.org/math/ap-statistics/xfb5d8e68:inference-quantitative-means/two-sample-t-test-means/v/example-of-hypotheses-for-paired-and-2-sample-t-tests

[35] Handbook of Biological Statistics http://www.biostathandbook.com/

[36] What Statistical Analysis Should I Use? https://stats.idre.ucla.edu/spss/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-spss/

[37] A simple decision chart for statistical tests. http://roslyndakin.com/wp-content/uploads/private/StatisticsDecisionTree.pdf

[38] Time Series Toolbox http://www.seec.uct.ac.za/time-series-analysis

[39] One-tailed and two-tailed tests https://stats.idre.ucla.edu/other/mult-pkg/faq/general/faq-what-are-the-differences-between-one-tailed-and-two-tailed-tests/

[40] Minimum Detectable Change and Power Analysis (EPA) https://www.epa.gov/sites/default/files/2015-10/documents/tech_memo_3_oct15.pdf

[41] TMDL Effectiveness Monitoring Planning Tool (EPA/ECY) https://www.epa.gov/sites/production/files/2015-09/emtool.xlsm

[42] Water Quality Portal (National Water Quality Monitoring Council)

https://www.waterqualitydata.us/

[43] Environmental Information Monitoring System (ECY)
https://apps.ecology.wa.gov/eim/search/default.aspx

[44] Dissolved Oxygen and Water Quality Standards (ECY)
https://apps.ecology.wa.gov/publications/documents/0210001.pdf

[45] OpenNSPECT
https://coast.noaa.gov/digitalcoast/tools/opennspect.html

[46] Tenmile watershed assessment, Whatcom Co. (NRCS NWQI)
https://www.whatcomcd.org/_files/ugd/74f091_f6499b3712af4cc38fda7c546d210163.pdf

[47] Perez, Michelle and Emily J. Cole. 2020. A Guide to Water Quality, Climate, Social, and Economic Outcomes
Estimation Tools: Quantifying Outcomes to Accelerate Farm Conservation Practice Adoption. Washington, DC: American Farmland Trust.
https://s30428.pcdn.co/wp-content/uploads/sites/2/2020/12/AFT_Outcomes_Tools_Web_4_21_21.pdf

[48] MacArthur, R. H. and E. O. Wilson. 1967. The Theory of Island Biogeography (Princeton University Press, Princeton, NJ).
https://www.pnas.org/doi/10.1073/pnas.1813051115

[49] Timmers et al. 2022. Conservation of Birds in Fragmented Landscapes Requires Protected Areas. Frontiers in Ecology and the Environment.
https://esajournals.onlinelibrary.wiley.com/doi/pdf/10.1002/fee.2485

[50] Wintle, B. A. et al. 2014. Global Synthesis of Conservation Studies Reveals the Importance of Small Habitat Patches for Biodiversity. Proceedings of the National Academy of Sciences.
https://www.pnas.org/doi/10.1073/pnas.1813051115

[51] Tscharntke, T., Klein, A. M., Kruess, A., Steffan-Dewenter, I., & Thies, C. (2005). Landscape perspectives on agricultural intensification and biodiversity–ecosystem service management. Ecology Letters, 8(8), 857-874.
http://www.bayceer.uni-bayreuth.de/mik/en/pub/pub/60936/EcolLetters2005,8_857-874.pdf

[52] Making Endangered Species Act Determinations of Effect for Individual or Grouped Actions at the Watershed Scale
https://www.oregon.gov/odot/GeoEnvironmental/Documents/Biology_NMFS_Endangered-Species-Determination.pdf

[53] High Resolution Change Detection (DFW)
https://hrcd-wdfw.hub.arcgis.com/

[54] National Land Cover Database (USGS)
https://www.usgs.gov/centers/eros/science/national-land-cover-database

[55] McGarigal, Kevin; Marks, Barbara J. 1995. FRAGSTATS: spatial pattern analysis program for quantifying landscape structure. Gen. Tech. Rep. PNW-GTR-351. Portland, OR: U.S. Department of Agriculture, Forest Service, Pacific Northwest Research Station. 122 p
https://www.fs.fed.us/pnw/pubs/pnw_gtr351.pdf

[56] Ament, R., R. Callahan, M. McClure, M. Reuling, and G. Tabor. 2014. Wildlife Connectivity:

Fundamentals for conservation action. Center for Large Landscape Conservation: Bozeman, Montana https://largelandscapes.org/wp-content/uploads/2019/05/Wildlife-Connectivity-Fundamentals-for-Conservation-Action.pdf

[57] McGarigal, K., SA Cushman, and E Ene. 2012. FRAGSTATS v4: Spatial Pattern Analysis Program for Categorical and Continuous Maps. Computer software program produced by the authors at the University of Massachusetts, Amherst. Available at the following web site: http://www.umass.edu/landeco/research/fragstats/fragstats.html

[58]The role of trees outside woods (TOWs) in contributing to the ecological connectivity and functioning of landscapes. https://www.woodlandtrust.org.uk/media/1719/role-of-trees-outside-woods.pdf

[59] Understanding Spatiotemporal Lags in Ecosystem Services to Improve Incentives https://academic.oup.com/bioscience/article/63/6/472/226229

[60] Sagelands Heritage Program https://www.conservationnw.org/our-work/habitat/sagelands/

[61] Protecting stream flows (ECY) https://ecology.wa.gov/Water-Shorelines/Water-supply/Protecting-stream-flows

[62] Monitoring and Evaluating Nonpoint Source Watershed Projects (EPA) https://www.epa.gov/nps/guidance-monitoring-and-evaluating-nonpoint-source-watershed-projects

[63] Monitoring Guidelines to Evaluate Effects of Forestry Activities on Streams in the Pacific Northwest and Alaska (EPA) https://digital.lib.washington.edu/researchworks/handle/1773/17068

[64] Insights for Developing Successful Agricultural Watershed Products (NRCS) https://content.ces.ncsu.edu/insights-for-developing-successful-agricultural-watershed-products

[65] Water Quality Monitoring for the Assessment of Watershed Projects (NRCS) https://content.ces.ncsu.edu/water-quality-monitoring-for-the-assessment-of-watershed-projects

[66] Quality assurance project plan examples (ECY) https://ecology.wa.gov/Issues-and-local-projects/Investing-in-communities/Scientific-services/Quality-assurance/Quality-assurance-for-NEP-grantees/Project-plan-examples

[67] The Volunteer Monitor's Guide to Quality Assurance Project Plans (EPA) https://www.epa.gov/sites/default/files/2015-06/documents/vol_qapp.pdf

[68] Edge-of-Field Water Quality Monitoring and Evaluation (NRCS) https://www.nrcs.usda.gov/Internet/FSE_DOCUMENTS/stelprdb1166615.pdf

[69] Quality-Assurance Plan for Water-Quality Activities in the US Geological Survey Washington Water Science Center (USGS) https://pubs.usgs.gov/of/2017/1044/ofr20171044.pdf

[70] Best Management Practices to Minimize Agricultural Phosphorous Impacts on Water Quality https://www.ars.usda.gov/is/np/bestmgmtpractices/best%20management%20practices.pdf

[71] Box and Whisker plots http://web.pdx.edu/~stipakb/download/PA551/boxplot.html

[72] Making Endangered Species Act Determinations of Effect for Individual or Grouped Actions at the Watershed Scale
https://www.oregon.gov/odot/GeoEnvironmental/Documents/Biology_NMFS_Endangered-Species-Determination.pdf