
Statistical Review of Wilderness Character Monitoring (WCM) for Gates of the Arctic National Park and Preserve

Prepared by:

Megan D. Higgs

PhD Statistics

Critical Inference LLC

1146 S. Pinecrest Dr.

Bozeman, MT 59715

406-640-1034

megan.higgs@critical-inference.com

Date March 1, 2021

Prepared for:

Gates of the Arctic National Park and Preserve

Main correspondence: Adrienne Lindholm, Grant Hilderbrand, and Kristin Pace

US National Park Service

Table of Contents

Review Summary	4
1 Introduction and Scope of Review	12
1.1 Wilderness character, measures, and thresholds for change	12
1.1.1 Definition of wilderness character and qualities	12
1.1.2 Measures to represent indicators of wilderness character qualities	13
1.1.3 Threshold for change associated with a measure	14
1.1.4 Equal weighting to obtain an “overall trend”	15
1.2 Unique aspects of GAAR and other Alaska parks	15
2 Review of statistics-related recommendations in WCM national guidance	16
2.1 Excerpts from Draft Technical Guide and GAAR WCMP	16
2.2 Defining or selecting measures	18
2.2.1 Summary and recommendations	19
2.3 Defining a threshold for meaningful change	20
2.3.1 Challenges and benefits associated with the <i>threshold for change</i>	20
2.3.2 Approaches to defining a <i>threshold for change</i>	21
2.3.3 Three options in national guidance	22
2.3.4 Summary and recommendations	25
2.4 Considering the role of statistical analysis	27
2.5 Use of linear regression	28
2.5.1 Regression models for assessing trend	28
2.5.2 Assumptions of the linear regression model	29
2.5.3 Summary and Recommendation	31
2.5.4 Threshold for change and the use of regression	31
2.5.5 Problems with shifting from percent change and statistically significant change	38
2.6 Statistical significance and default alpha levels	39
2.6.1 Statistical hypothesis tests and decision making	40
2.6.2 Alpha-levels, Type-I errors, and use of p-values	40
2.6.3 A default type-I error rate regardless of measure	41
2.7 Making the reference for judging change clear	43
2.8 Statistical wording (confidence, certainty, alpha level)	44
2.9 Cautions related to automation	46
2.10 Summary and recommendations	47
3 GAAR-Specific Considerations and Recommendations	49

3.1 All Thresholds for Change options included in Draft GAAR report	49
3.2 Natural Quality – Plants from Draft GAAR WCMP	51
3.3 Challenges for GAAR.....	52
3.3.1 Data and sampling	53
3.3.2 Pristine baseline of no non-native plant species.....	53
3.4 Measures proposed for Plants in WCM documentation	54
3.4.1 Index of Invasive Plant Species.....	54
3.4.2 Count of total non-native species.....	56
3.5 Alternative approaches	57
3.5.1 Categories defined according to meaningful change	57
3.5.2 A multi-stage or adaptive approach	58
3.5.3 Beyond invasive species.....	59
3.6 Potentially useful questions for developing the GAAR WCMP	59
3.7 Additional recommendations and considerations	60
3.7.1 <i>Plants</i> -specific recommendations.....	60
3.7.2 General recommendations.....	60
References	63
APPENDIX A – Supporting Resources and Documentation	65
A.1 Assessing quantitative properties of measures for a WCMP	65
A.2 Percent change and additive change	69
A.3 Meaningful change backdrop	70
A.4 Translate p-value threshold	71
APPENDIX B – Original Scope of Work (SOW)	74

Review Summary

Wilderness character is a holistic concept based on the interaction of (1) biophysical environments primarily free from modern human manipulation and impact, (2) personal experiences in natural environments relatively free from the encumbrances and signs of modern society, and (3) symbolic meanings of humility, restraint, and interdependence that inspire human connection with nature. Taken together, these tangible and intangible values define wilderness character and distinguish wilderness from all other lands.

Keeping It Wild 2 (2015)

Overview and Scope of the Review

The Wilderness Character Monitoring program is designed to provide a foundation for developing and implementing a nationally consistent approach to monitoring wilderness character. Four agencies manage wilderness lands: National Park Service, Bureau of Land Management, Fish and Wildlife Service, and Forest Service. The Wilderness Character Monitoring program is described in *Keeping It Wild 2: An Updated Interagency Strategy to Monitor Trends in Wilderness Character Across the National Wilderness Preservation System* (KIW2 2015). For the National Park Service, KIW2 is accompanied by a *Draft NPS Wilderness Character Monitoring Technical Guide* containing general guidance and specific recommendations at the national level; the November 20, 2020 pdf version is hereafter referred to as the *Draft Technical Guide*.

The National Park Service (NPS) in Alaska is in the process of developing Wilderness Character Monitoring Plans (WCMPs) for its parks and this biometric review was solicited as part of that process, with an emphasis on Gates of the Arctic National Park and Preserve (GAAR). This review evaluates components of the WCM process from a statistical perspective. It is not a thorough review of the entire process. The scope covers two main goals: (1) provide feedback regarding recommendations in the *Draft Technical Guide* that rely heavily on statistical concepts and methods, and (2) assist GAAR in the process of selecting and justifying a measure and associated *threshold for change* for the Natural Quality *Plants* indicator. The first goal applies broadly to any park planning to rely on the *Draft Technical Guide* (note that many recommendations in the *Draft Technical Guide* come from KIW2). The second goal highlights challenges GAAR and other Alaska parks may face in developing and implementing WCMPs using national guidance.

This *Review Summary* provides an overview of conclusions and recommendations based on the review, with supporting details provided in the text of the report.

I. General conclusions and recommendations

The WCM process uses mainly quantitative measures to make conclusions regarding trends in pre-specified indicators of five qualities of wilderness character (KIW2 2015). This is inherently a difficult task given the qualitative social and human aspects defining wilderness character (see opening quote). Current national guidance relies on statistical methodology for estimating trends and setting criteria for judging whether observed trends represent meaningful change in wilderness character. The *Draft Technical Guide* raises multiple concerns from a statistical perspective, particularly related to the *threshold for change*, which is crucial to assigning the trend categories (“declining”, “stable”, or “improving”) that are the basis of conclusions and reporting for the WCM process. This review focuses

on statistical aspects of the process that are explicitly described in the national documents; there are other very important measure- and park-specific statistical issues, such as those related to sources of variability and how data are collected over time and space.

A threshold for change is defined by the question: *What amount of change in a measure qualifies as meaningful relative to wilderness character?* Justifying an answer to this question is challenging, and the choice ultimately requires human judgment and qualitative reasoning; there is no single or correct “answer.” The choice should depend on knowledge of past, current, and future conditions of the wilderness area and an understanding of how associated conclusions may influence decision-making; as well as more technical considerations regarding properties of the indicator and proposed measure(s) and how data used for the measure are collected.

This eight-page *Summary* provides the main findings of the review and associated recommendations, starting with general comments applicable on the national level and ending with comments specific to GAAR. Recommendations generally encourage moving away from overly prescribed or default methods in favor of approaches honoring the challenges involved in monitoring the complex idea of wilderness character with a relatively small set of quantitative measures. A prescriptive process is attractive in the sense that it provides consistency across parks; is easier to implement; appears less affected by individual human judgments and biases; and requires less justification by individual parks. However, placing high value on making the process prescriptive comes at a cost – it can lead to an unintended loss of connection between the goal of monitoring changes in wilderness character and the measures (and *thresholds for change*) ultimately used as the basis for conclusions. In particular, statistical methods and default criteria should not be substituted for context-dependent justification of decisions.

Topic: Threshold for Change

Summary A *threshold for change* chosen according to the options presented in the *Draft Technical Guide* (pages 19-22) may fail to represent meaningful change in an indicator of wilderness character. Conclusions regarding trend depend on choice of threshold, making it a crucial part of the WCM process.

Recommendations:

- The step of defining meaningful change should rely on professional expertise and context for justification and precede statistical analysis. Statistical tests and p-value cutoffs are not designed to answer the question of what amount should be judged as “meaningful.”
- Rigor can be improved by setting an expectation for justifying the choice of *threshold for change* to a broad audience on the scale of the chosen measure and in a way that directly aligns with the definition of wilderness character.

Summary The *Draft Technical Guide* emphasizes the importance of “measure selection,” but downplays the importance of assigning associated thresholds.

Recommendations

- Defining a *threshold for change* should be explicitly included in Step 2 of the WCM process (pages 16-17 *Draft Technical Guide*).
- The reasons for choosing a particular *threshold for change* should be transparent and explicitly justified relative to the measure and the ultimate decision about trend, rather than presented as if it is more of a technical aspect of the analysis process.
- Usefulness of a conclusion regarding trend in an indicator of wilderness character depends on the interplay between the measure and the threshold for change. Therefore, choice of measure and *threshold for change* should be considered simultaneously.

Summary The three options for *threshold for change* presented in the *Draft Technical Guide* (“any change,” “percent change,” or “Statistically significant change based on regression”) are overly restrictive and do not encourage context specific choices.

Recommendations

- The *Draft Technical Guide* describes flexibility in choice of measure afforded to each NPS wilderness area; that flexibility should also apply to defining *thresholds for change* given that conclusions will be sensitive to this part of the process. In general, terms such as “strongly encouraged measures” and “three options for defining threshold for change” may discourage deviation from the recommendations even when justified.
- Consider including other options in the *Draft Technical Guide*, such as “additive change,” as well as encouraging creativity in developing categorical measures and associated thresholds simultaneously.
- Basing a *threshold for change* on “statistically significant change based on regression” is not recommended for use as a default.
- Build in an expectation to spend more effort in the planning phase to develop park-specific measures *and* thresholds, with the goal of building a strong foundation leading to well-supported conclusions with less need for justification later in the process.

Topic: Default use of regression and statistical significance

Summary Using “statistically significant change based on regression” as the method for setting a *threshold for change* has severe limitations in its usefulness and defensibility relative to the goals of WCM. The preference given to this option seems to rely on an unstated assumption that, if at all possible, regression and statistical significance should be used because it provides the most rigorous approach available. This assumption is not justified or defensible.

Recommendations

- Remove statistical hypothesis testing as a default approach for defining a *threshold for change*. In general, use of statistical methods does not automatically increase rigor of the WCM approach, as implied in the national documents. Relying on statistical methods and commonly used cutoffs is often viewed as more scientifically rigorous than relying on approaches perceived as less complex or sophisticated; however, “rigor” relies on how well the model is aligned with the practical questions and goals, and the extent to which model assumptions are justified.
- Revise wording that implies the use of statistical methodology renders an approach more rigorous and/or useful than other options. Statistical approaches often carry a perception of “objectivity” and lack of “bias” that is rarely questioned. It can be attractive to rely on such an approach, particularly when it makes the process more automatic and easier to implement -- not only in carrying out the calculations, but because it often comes with fewer expectations in terms of having to justify one’s approach.
- Revise specific wording related to statistical methods that may be misleading or reflect misinterpretations. For example, complexity of a measure and availability of at least five measure values are not well justified reasons for use of linear regression, and “statistical rigor” is not gained simply by using a lower alpha-level.
- The process for specifying a *threshold for change* should stay aligned with the goal of defining and justifying what amount of change in a measure represents meaningful change in wilderness character; a statistical model and test do not automatically provide that information.
- Instead of placing preference on a statistical hypothesis test, it is recommended that preference be placed on approaches explicitly justified and able to be critiqued by stakeholders.

Topic: Lack of transparency to users and stakeholders

Summary The use of statistical significance disconnects the *threshold for change* from a scale and context that can be broadly evaluated and understood. The option discourages meaningful review by stakeholders who lack statistical expertise, instead putting them in a position to blindly trust conclusions. Difficulty in explaining a threshold broadly to stakeholders in a way that can be easily evaluated should be taken as a potential red flag regarding the usefulness of the conclusions.

Recommendations

- If regression is used, the p-value threshold (applied to a particular collection of observations) can be translated into a slope of a line - interpreted as the additive change in the mean of the measure per year over the time period. If regression and/or statistical significance is used, it should be accompanied by explicit translation of the threshold into language understandable by a broad audience, rather than hidden as a p-value cutoff.
- The regression approach does not provide a direct comparison to the baseline measure as described as a goal in the *Draft Technical Guide*; a different approach should be selected if stakeholders are under the impression that conclusions are relative to a baseline and/or this characteristic of using regression should be explicitly described.

Summary A threshold based on “statistically significant” looks on the surface as if it is not changing over time (e.g., p-value cutoff remains at 0.10), but the magnitude of change considered meaningful *does* change over time as new observations are added.

Recommendations

- The *threshold for change* specified for a measure should remain consistent over time, unless a change is explicitly justified; this advice is consistent with the recommendation in the *Draft Technical Guide*. However, this advice runs counter to use of “statistically significant change based on regression.” Resolving this inconsistency will help users and other stakeholders.

Summary The *Draft Technical Guide* states “The status of wilderness character in the baseline assessment is the reference point against which change in wilderness character is assessed and evaluated over time.” This statement is inconsistent with other wording and recommendations and results in a lack of clarity regarding the definition of “trend.”

Recommendations

- Current guidance is unclear about whether the “baseline measure value” will continue to be used as a reference point after more data are collected, as opposed to the previous time period’s value or a summary of values over multiple years (e.g., 3-year average). Improving clarity in recommendations and/or acknowledging the need for flexibility is recommended.
- Explicit definition of “trend” should include detail about reference points, time periods, and natural annual variability. More attention to attempting to define “trend” for specific types of measures in the context of WCM may encourage consistency over time and across wilderness areas. As is, conclusions regarding trend may differ substantially depending on what reference and summary are used (e.g., direct comparison to baseline value vs. linear regression). This aspect of the process is also crucial to defining a relevant *threshold for change*, choosing a method to estimate change, and consistently implementing the proposed process after a threshold is specified.
- Ensure that what will be implemented by the WCM database (WCMD) after measure values are input matches what users expect will happen. For example, is WCMD comparing to baseline when users expect it to compare to the previous measure value, or vice-versa?

Topic: Defining and Selecting Measures

Summary WCM does not intend to plan or implement new sampling efforts to collect data for measures; instead, data collected for other monitoring efforts or objectives are to be used to avoid duplication and encourage efficient use of resources. This restricts the possible options for measures for each indicator, which can introduce additional challenges and limitations in terms of aligning measures with specific WCM indicators (i.e., how well does a measure capture what it's meant to measure?).

Recommendations

- This approach may be necessary for logistical and cost reasons; however, the limitation introduced by the restriction can be explicitly discussed in the WCMPs. Types of limitations include assessing how well the chosen measure actually captures information about the indicator it's meant to measure, and addressing issues related to data collection and scope of inferences (e.g., difficulties in separating changes over space from changes over time).
- As a general strategy, it is recommended that parks first identify what measure they would use *if* resources for data collection were unlimited. That is, rather than starting from the list of available data and measures, start the process from deep consideration of what would be ideal (or at least closer to ideal). Then, work to keep as many of the characteristics of the "ideal" measure as possible when constraining the final measure based on available data and resources. Documenting differences between the ideal measure and chosen measure may help in describing and understanding limitations and challenges relative to wilderness character.

II. Alaska- and GAAR-specific summaries and recommendations

Gates of the Arctic National Park and Preserve (GAAR) consists of 8.4 million acres in Alaska, all north of the Arctic Circle. More than 7 million acres are designated wilderness and an additional 914,000 acres within GAAR are eligible for wilderness designation. The GAAR Wilderness borders the designated wilderness of Noatak National Preserve, and the two areas together make up the largest contiguous area of designated wilderness in the National Park System (Draft GAAR WCMP, 2020).

Vast Alaskan wildernesses introduce unique challenges in terms of adopting national-level recommendations, such as those in the *Draft Technical Guide*. Baseline wilderness character and potential scenarios for change in Alaska may be very different from those found in the lower 48 states, as illustrated through considering the *Plants* indicator of the Natural Quality of wilderness character. The *Plants* example also highlights the importance of considering choice of measure and *threshold for change* concurrently.

Recommendations and considerations for GAAR

- Wilderness character is inherently a qualitative concept (see definition in opening quote) and difficult to measure, particularly when restricted to quantitative metrics based on available data. It is possible to use the currently proposed WCM process to develop measures and thresholds for change that more holistically capture indicators of wilderness character. For example, categories can be defined based on multiple criteria (including qualitative information) such that a shift in category represents meaningful change in wilderness character. This idea is actually referred to in the example briefly referenced within the description of the "any change" threshold in the *Draft Technical Guide*.
- Alaska parks are encouraged to take advantage of flexibility in the process afforded to individual parks to define measures *and* thresholds for change tailored to the ecosystem and management contexts. Development of area-specific measures and thresholds will require more work in the

creation of the WCMP, but less work and justification later to support conclusions and potential management decisions.

- As further discussed in the next section, the national guidance defaults to an emphasis on infestation of invasive species; an emphasis that makes sense for the lower 48 states. For Alaska, *other potential changes in plant communities may be of higher priority relative to wilderness character*, such as changes in the distribution of native species over the landscape in (e.g., increase in shrubs and spruce growing in areas previously tundra). Shifts in the spatial distribution of species are expected to continue due to climate change and the strongly encouraged measure focused on monitoring non-native or invasive plant species would miss such change.
 - A possible measure suggested in discussion with Peter Neitlich is % area covered by shrubs (advancement of woody vegetation), along with a quantitative threshold such as a change in cover by 1%.
 - Measures to capture relative changes among native species need not be restricted to only typical quantitative summaries. For example, it could be as simple as shrubs becoming the main cover in a particular area and/or spruce being found in a particular area of tundra.
 - With continued climate change, trends in such measures are expected to be “stable” or “declining” but not improve; this may be important relative to the usefulness of the WCM process for management decisions.
- If a measure based on non-native plants will be used, it may be *worth considering a multi-stage approach* where one measure is used until non-native species are detected (e.g., presence/absence of species of potential concern), followed by a shift to a measure accounting for potential and observed impact after infestation in a way that makes sense for Alaska. The definition of the second stage measure(s) and threshold(s) could be initiated after pre-specified criteria are met (e.g., detection of a particular non-native species) and depend on the state of the system at the time (e.g., depending on what and where species are found, or suspected to be found in the near future).

Potentially useful questions to consider as WCMP is developed:

- What is considered the biggest threat to wilderness character in terms of the *Plants* indicator for GAAR?
- What are the potential negative aspects of missing or ignoring information relative to changes in spatial distribution and extent of native species?
- What are the potential negative aspects of missing or ignoring information about non-native or invasive species?
- Is it worth using two different measures (one for invasive species and one for distribution of native species)? Note: If two measures are created, the default is to give them equal weight in the WCM process, and therefore this should be considered in the process of measure and threshold development.
- Are there other aspects related to the *Plants* indicator that should also be considered for GAAR wilderness character?
- What data are or will be available to support each potential measure?
- What process will be used to develop and justify a *threshold for change* simultaneously with each measure?
- How might resulting conclusions regarding trend (based on the measure and threshold), potentially inform management or policy decisions? How sensitive should the threshold be to change? Should it serve as a red flag for coming change or should it convey change that has clearly happened already?

Previously proposed measures

The strongly recommended measure for the *Plants* indicator provided in the *Draft Technical Guide* (pages 42-46) is an *Index of Invasive Plant Species* (the *Index*). General characteristics of the measure are briefly summarized here to provide context for GAAR-specific considerations.

Calculating the Index of Invasive Plant Species score:

- The *Index* combines information about (1) presence/absence (list of detected invasive species); (2) species-specific potential impact using I-ranks (US Invasive Species Impact Ranks conveyed through integer scores: 0, 1, 2, or 3); (3) species-specific observed impact (overall trends in total area infested summarized by an integer score: 0, 1, 2, or 3); and (4) whether a species co-occurs with a federally listed animal or plant species (0 or 1).
- For species with trend in infestation classified as “declining” or “unknown,” the minimum score is 1 if an I-rank is available for the species, and 0 if an I-rank is not available. The maximum score for a species is 7, corresponding to an I-rank of 3 (“high”), infestation trend of 3 (“increasing rapidly”), and co-occurrence with a federally listed species.
- The overall index score for an area is obtained by summing the scores from all species on the list.

Interpreting change in the Index score:

- The *threshold for change* is described by “any change in the index value is considered significant.”
- A change in the value of the *Index* reflects a change in the number of invasive species detected *and/or* in the trend of infestation for species already present (spreading faster or slower). The “any change” threshold does not distinguish between the two types of change in terms of a conclusion about trend.
- Conclusions about trend do not take the magnitude of the baseline (or previous) value into account. For example, changes from 0 to 1, 0 to 7, and 15 to 16 would all result in “declining trend,” despite potentially representing very different practical scenarios. This illustrates the importance of considering the measure and the threshold together. In this case, the *Index* includes valuable information about potential and observed impacts, but when used with the “any change” threshold the depth of information is not reflected in the conclusion about trend.
- The *Draft Technical Guide* is unclear about whether the reference for assessing change is meant to always be the baseline value, or if it is expected to shift over time. For example, if over the first 5 years the index changes from 0 to 10 (“declining trend”) and then decreases from 10 to 4 in the subsequent 5-year period, does that deserve a “declining trend” (relative to baseline) or an “improving trend” (relative to the previous time period), or is there a recommendation for conveying both?

Evaluating use of the Index of Invasive Species for GAAR

Advantages include:

- Consistency with national guidance and other parks
- Serves to detect, as meaningful change, the shift from no known invasive plant species to the detection of at least one species (assuming information is available for I-rank(s))
- When/if an invasive species is detected, the *Index* score provides information about the potential severity of infestation (assuming an I-rank is available for the species and that it applies to the GAAR ecosystem).
- May capture information about human related spread of invasive species (e.g., airstrips, ATVS, backpackers, etc.).

Disadvantages include:

- The information beyond presence/absence is not immediately useful given GAAR’s baseline of no known invasive species.

- The “any change” threshold may document detection of an invasive species, but will not be sensitive to what species is found.
- The *Index* uses a summary of trend in amount of wilderness area infested, which may be difficult to obtain given the vast areas of AK parks, relatively small areas that would be initially occupied, and constraints on amount of area that can be sampled. This part of the measure would likely require additional definitions and guidance to make it useful for AK.
- If a detected species has an unknown I-rank and unknown trend in infestation, the *Index* value will be zero and fail to record it as a change by the stated criterion. The extent of this problem depends on availability of useful I-ranks and extent of infestation data, both of which may present greater challenges in Alaska than the lower 48.
- If values are always compared to the baseline measure value of zero, it will never be possible for GAAR to have an “improving trend” conclusion. This is not necessarily a disadvantage, but should be considered relative to the WCM process and management decisions.
- The *Index* only includes species classified as invasive (potential for economic or environmental harm) rather than all non-native species. Alternatively, this can be an advantage if the park does not consider all non-native species to be a potential threat to wilderness character.
- The *Index* does not account for weedy native species whose distribution over space may change in response to climate change, potentially affecting wilderness character. More generally, it does not account for changes in the distribution of vegetation cover classes over the landscape (e.g., increase cover by shrubs or spruce trees moving into areas that were previously tundra).

The measure and its threshold currently described in the Draft GAAR WCMP is the **count of the total non-native species with an “any change” threshold**, which shares many characteristics with the *Index*, particularly when considered with the “any change” threshold.

Advantages include:

- Ability to capture a shift from no known species to presence of non-native species
- Includes all non-native species, in addition to invasive species (the importance of this property depends on how the additional species are connected to wilderness character)
- An I-rank is not needed for each species
- A summary of trend in the area of the wilderness infested is not needed
- Will be greater than zero if any non-native species are detected, thus avoiding the potential problem of the *Index* remaining zero if a detected species has an unknown I-rank

Disadvantages include:

- Does not include the species-specific information potentially captured in the *Index*, such as potential for and observed extent of infestation. Instead, all non-native species are treated equally. However, when both are used with “any change” there may be little difference between the two measures in terms of conclusions about trend in wilderness character.
- As with the *Index*, if the reference is always the baseline measure value, then “improving trend” will never be an option in GAAR (this may not be a disadvantage).
- As with the *Index*, it does not capture infestations of weedy native species or changes in spatial distribution of other species (e.g., an increase in shrubs or spread of spruce into tundra).

1 Introduction and Scope of Review

The Wilderness Character Monitoring program is designed to provide a foundation for developing and implementing a nationally consistent approach to monitoring wilderness character. Four agencies manage wilderness lands: National Park Service, Bureau of Land Management, Fish and Wildlife Service, and Forest Service. The Wilderness Character Monitoring program is described in *Keeping It Wild 2: An Updated Interagency Strategy to Monitor Trends in Wilderness Character Across the National Wilderness Preservation System* (KIW2 2015). For the National Park Service, KIW2 is accompanied by a *Draft NPS Wilderness Character Monitoring Technical Guide* containing general guidance and specific recommendations at the national level (the November 20, 2020 pdf version is hereafter referred to as the *Draft Technical Guide*).

The National Park Service (NPS) in Alaska is in the process of developing Wilderness Character Monitoring Plans (WCMPs) for its parks and this biometric review was solicited as part of that process, with an emphasis on Gates of the Arctic National Park and Preserve (GAAR). This review evaluates components of the WCM process from a statistical perspective, focusing on language and recommendations found in current versions of national guidance; it is not a thorough statistical review of the entire process. The scope covers two main goals: (1) provide feedback regarding recommendations in the *Draft Technical Guide* that rely heavily on statistical concepts and methods, and (2) assist GAAR in the process of selecting and justifying a measure and associated *threshold for change* for the Natural Quality *Plants* indicator. The first goal applies broadly to any park planning to rely on the *Draft Technical Guide* (note that many recommendations in the *Draft Technical Guide* come from KIW2). The second goal highlights challenges GAAR and other Alaska parks may face in developing and implementing WCMPs using national guidance. This technical report is organized into two main sections (Section 2 and Section 3) corresponding to the two goals.

1.1 Wilderness character, measures, and thresholds for change

The WCM process uses mainly quantitative measures to make conclusions regarding trends in pre-specified indicators of five qualities of wilderness character (KIW2 2015). This is inherently a difficult task given the qualitative social and human aspects defining wilderness character.

Current national guidance relies on statistical methodology for estimating trends and setting criteria for judging whether observed trends represent meaningful change in wilderness character. The *Draft Technical Guide* raises multiple concerns from a statistical perspective, particularly related to the *threshold for change*, which is crucial to assigning the trend categories (“declining”, “stable”, or “improving”) which are the basis of conclusions and reporting for the WCM process. This review focuses on statistical aspects of the process that are explicitly described in the national documents; there are other very important measure- and park-specific statistical issues, such as those related to sources of variability and how data are collected over time and space.

1.1.1 Definition of wilderness character and qualities

A definition of “wilderness character” is provided in KIW2 (2015) that conveys the complexity of the concept and the associated difficulties in measuring and monitoring:

Wilderness character is a holistic concept based on the interaction of (1) biophysical environments primarily free from modern human manipulation and impact, (2) personal experiences in natural environments relatively free from the encumbrances and signs of modern society, and (3) symbolic meanings of humility, restraint, and interdependence that inspire human connection with nature. Taken together, these tangible and intangible values define wilderness character and distinguish wilderness from all other lands.

Keeping It Wild 2 (2015)

The five qualities of wilderness character monitored are *Untrammeled, Natural, Undeveloped, Solitude or Primitive and Unconfined Recreation, and Other Features of Value*. Indicators have been identified for each quality and then measures are defined to capture potential change in the indicators over time (see list on pages 13-15 in the *Draft Technical Guide*). The above definition and list of qualities is a reminder of the challenges in measuring wilderness character, in classifying trend neatly into categories, and ultimately in combining information across all the qualities. Similar challenges arise in adaptive monitoring and the perspective taken in this review is in line with that stewardship and this review takes a similar perspective to that described in Cundill et al. (2011).

1.1.2 Measures to represent indicators of wilderness character qualities

The second step in implementing the WCM Process in the NPS is **Select measures and develop a Wilderness Character Baseline Assessment** (page 16 of the *Draft Technical Guide*). The selection of reasonable measures is a crucial part of the WCM process; conclusions will ultimately depend on how well the measures represent the indicators of the wilderness character qualities, as well as how the measures are analyzed to arrive at a conclusion regarding trend.

Ultimately, flexibility is afforded to individual parks to choose relevant measures and justify their choices. However, the *Draft Technical Guide* also “strongly encourages” the use of certain measures, which may create some tension in terms of the direction a park may decide to go. On page 13, arguments are provided for using the strongly encouraged measures, whereas elsewhere the wording acknowledges park differences and need for change. For example, the following wording is provided for the basic step of *Select Measures* (pg 16):

Select at least one measure for each indicator of each tangible wilderness quality to describe its current condition and track change over time. Parks should first consider the strongly encouraged measures in this guide and determine if the measures are appropriate for that wilderness area. Determining the appropriateness of measures in this guide and selection of any additional measures is a collaborative project between park staff, with assistance from regional/national support offices as needed.

The choice of measures is also described as data-dependent because WCM is designed to be implemented using data collected for other monitoring or data collection efforts. This constraints this places on measure selection may be unavoidable, but it is worth acknowledgment and discussion in terms of implications it may have on the ultimate conclusions. *Draft Technical Guide* describes the potential role of available data on pages 12-13:

All the measures included in this Guide were developed to reduce the amount of time and effort required of parks to implement WCM, such that existing data are used whenever appropriate and available. Parks, however, may need to utilize datasets that are not listed in this protocol to collect the requisite data or compile new data from the field for some measures (e.g., tracking the number of authorized trampling actions, or assessing the condition of unique features integral to wilderness character). Whenever possible, protocols were developed to use data from national or regional monitoring programs (e.g., air quality, water quality). For some measures, national staff compile and provide data to the park for verification. For other measures, park staff compile data from existing databases, administrative records (e.g., minimum requirement decisions), professional knowledge and judgment, or field collection.

Given the measure and park- specific aspects regarding data to support each measure, this part of the process is not covered in depth in this review, despite its importance to the overall process. For example, for the purpose of this review, it is assumed that sampling plans are being carefully evaluated relative to objectives and limitations based on constraints in data collection are acknowledged and considered in the new setting of WCM. Such aspects are included in the assessment of data quality in the WCM process. General sampling-related recommendations are beyond the scope of this review and should be considered on a case-by-case basis, though Appendix A.2 of this report provides a start to some guiding questions in relation to how data are collected over time and space.

The measure selection step in the NPS WCM process does not mention assignment of a *threshold for change*, though it may be implied because it is a necessary input to the Wilderness Character Monitoring Database (WCMD) in order for the trend conclusions to be made (applying the criteria). Measure selection is clearly crucial to the process, but the importance of the *threshold for change* may be under-appreciated, or at least not conveyed as crucial in the national guidance. The decision process regarding classification of trends as “stable,” “declining,” or “improving” depends on the interplay between the measure and its assigned threshold for change, thus considering them together can improve the usefulness of conclusions.

1.1.3 Threshold for change associated with a measure

A *threshold for change* is defined by the question: *What amount of change in a measure qualifies as meaningful relative to wilderness character?* Justifying an answer to this question is challenging, and the choice ultimately requires human judgment and qualitative reasoning; there is no single or correct “answer.” The choice should depend on knowledge of past, current, and future conditions of the wilderness area and an understanding of how associated conclusions may influence decision-making, as well as more technical considerations regarding properties of the proposed measure(s) and associated data. Adopting default thresholds or methods for setting thresholds may introduce limitations in terms of the usefulness of conclusions and the ability to explain and justify them to stakeholders.

As mentioned above, the process of selecting a measure should not be separate from the process of defining a *threshold for change* associated with that measure; the usefulness and relevancy of a measure will depend on the ability to set a meaningful threshold for change. The connection between these two parts of the process is referred to throughout the report and much of the statistical review revolves around statistical language and recommendations provided in the context of setting a *threshold for change*.

1.1.4 Equal weighting to obtain an “overall trend”

The WCM process ultimately combines the trend classifications obtained for individual measures (by applying their associated thresholds for change) across measures, indicators, and qualities to get an overall trend classification for a park. It is important to realize that the default way of proceeding through the process (as implemented in the WCMD) is to equally weight measures, indicators, and qualities when their information is combined into an “overall trend of wilderness character” (see pages 24-29 of *KIW2*). This quantitative aspect of the process is not covered in detail within this review, but is something to take into account. Equal weighting is often assumed to be a default “no weighting,” but it is in fact a choice made that affects conclusions. If the equal weighting is a constraint that must be kept in the WCMD, then choice of measures and thresholds for change should try to take into account how the information will be combined later in the process so that equal weights are as reasonable as possible.

1.2 Unique aspects of GAAR and other Alaska parks

Gates of the Arctic National Park and Preserve (GAAR) consists of 8.4 million acres in Alaska, all north of the Arctic Circle. More than 7 million acres are designated wilderness and an additional 914,000 acres within GAAR are eligible for wilderness designation. The GAAR Wilderness borders the designated wilderness of Noatak National Preserve, and the two areas together make up the largest contiguous area of designated wilderness in the National Park System (Draft GAAR WCMP, 2020).

Vast, remote, and pristine Alaskan wildernesses introduce unique challenges in terms of adopting national-level recommendations, such as those in the *Draft Technical Guide* and *KIW2*. Baseline wilderness character and potential scenarios for change in Alaska may be very different from those found in the lower 48 states, as illustrated through consideration of the *Plants* indicator of the Natural Quality of wilderness character in Section 3 of this report. The *Plants* example also highlights the importance of considering choice of measure and *threshold for change* concurrently.

2 Review of statistics-related recommendations in WCM national guidance

This review focuses on the *Threshold for Change* and *Protocol* sections in the *Draft Technical Guide*, although information from other sections is integrally connected and also important to the overall WCM process. Ultimately the process is designed to assign trend categories (stable, upward, or downward) to each measure, which are then combined to produce higher level trend conclusions for indicators, qualities, and then overall for a park. The usefulness of the “overall” trend conclusion of the WCM process for a time period and a park depends on justification of all the included components; the most basic level is the selection of each measure and definition of meaningful change in that measure (*threshold for change*), as well as the quality and quantity of information going into the measure.

This section identifies areas and recommendations found in the *Draft Technical Guide* that are based on or refer to statistical methods or reasoning. The information in this section applies beyond GAAR to any park relying on the *Draft Technical Guide* to carry out the WCM process. And, much of the information in the NPS *Draft Technical Guide* comes from the broader guidance in *Keeping It Wild 2*.

Section 2.1 contains verbatim excerpts from the *Draft Technical Guide* which are the basis of discussion in this section. The excerpts are provided for ease of reference and to provide a summary of the type of statements and information on which this review is focused.

2.1 Excerpts from Draft Technical Guide and GAAR WCMP

Excerpts from the *Draft Technical Guide* relevant to this review are included here for ease of reference. They include parts of the *Sections Described for Each Measure* and *Details About Data Adequacy and Determining Trend* section, as well as a few definitions. Bold face has been added to call attention to the most relevant parts of the excerpts.

[pg 19-20 *Draft Technical Guide*, pg 14 *Draft GAAR WCMP*; bold face added]:

Threshold for change. This section identifies the **amount of change necessary to qualify as a meaningful change in the measure**. This threshold varies by measures based on: 1) inherent temporal variability of the data for some measures; and 2) the adequacy of the data.

- Three standard categories are used for thresholds:
 - Any change
 - The any-change threshold applies to measures for which any change in the data would be meaningful from the perspective of wilderness character. This threshold is typically used for measures where change over time is unlikely and where there is high certainty about changes in the data.
 - For example, this threshold applies to the measure Acres of Inholdings [link] because change over time is relatively infrequent, and any increase or decrease in inholding acres would be a meaningful change. The any-change threshold is also used for measures where categories are used to determine the trend in the measure. For example, when professional judgment is used for the measure Index of Invasive Plant Species [link], the any-change threshold is used because any change between the defined categories (none, low-, moderate-, or high-

estimated percentage of a wilderness occupied by nonindigenous plants) would be, by definition, a meaningful change.

- Percent change
 - The percent-change threshold applies to measures that are less sensitive to change, that show variation from year to year, or that have medium data adequacy. Three types of percent-change thresholds are assigned to measures: 3-percent change, 5-percent change, and 10-percent change. The larger percentages indicate a higher likelihood of annual variation or a lower expected data adequacy.
- Statistically significant change based on regression
 - The regression threshold applies to certain measures once they have accumulated a sufficient amount of data, and is used because it provides statistical rigor in the long-term analysis of trends in the measures. Using regression to determine whether a trend is statistically significant requires the user to assign the desired degree of confidence, or certainty, in the results, called the *alpha* level (there is an extensive literature on this topic that is beyond the scope of this technical guide). For all measures that use regression in this guide, an *alpha* level of 0.1 will be used in determining statistical significance, meaning that there is a 10-percent chance of concluding that there is a significant trend when in reality there is not a trend, or conversely that there is 90-percent confidence or certainty that the trend is real. This *alpha* level allows an appropriate balance between the need to catch trends early while maintaining as much statistical rigor as possible in correctly identifying meaningful trends (see Appendix B in Landres et al. 2009 for details on the selection of this *alpha* level and use of regression).

[pg 18-19 *Draft Technical Guide*, pg 13-14 *Draft GAAR WCMP*; bold face added]:

Frequency. Identifies two frequency components per measure:

- 1) Frequency of data collection (or compilation) Per guidance in KIW2, WCM measures should be entered into the WCMD **at least once every five years**. Some measures only need data collection/compilation and reporting at 5-year intervals because the data are unlikely to change during this period. Other measures, however, will require more frequent data compilation (i.e., in years between entry into the WCMD) because the data are likely to change multiple times over a five-year period and need to be assessed more frequently.
- 2) Frequency of data entry in the WCMD. For data entry in the WCMD, **entries should reflect a measure value appropriate to compare against the baseline assessment value for purposes of determining trend.**

[pg 22 *Draft Technical Guide*, pg 18 *Draft GAAR WCMP*; bold face added]:

Determining Trend in overall wilderness character

WCM supports managers in assessing the current status of wilderness character in parks. Assessing overall trend in wilderness character provides a readily interpretable assessment that can help inform future management actions to better preserve wilderness character over time.

This trend is determined using nationally consistent rules to compile trends across the measures, indicators, monitoring questions, and qualities to derive an overall trend in wilderness character for every wilderness area. A trend of stable, upward, or downward is derived for each measure based on the data values and the corresponding threshold for

significant change. Trends, not the data, in each measure are compiled by using rules to identify the trend in the indicator. More information about these rules are described on pages 24-29 of KIW2.

[pg 22 *Draft Technical Guide*, pg 18-19 *Draft GAAR WCM*; bold face added]:

Determining significant change and trend in a measure

A significant, or meaningful, change in each measure is based on the “Threshold for Change” section for each measure. In a given year that trend is reported, **measures that have at least five data points, simple linear regression should be used to determine trend. For measures where the five-year minimum monitoring interval yields only one measure value, the “trend” in a measure is determined by comparing the most recent measure value (most commonly a three-year rolling average) with the earliest available baseline data for that measure. While this trend is not as robust as the linear regression analysis,** it will give managers a preliminary sense of measure trend until such time that five data points are available.

[pg 18 *Draft GAAR WCM*; shorter version on pg. 10-11 of *Draft Technical Guide*; bold face added]:

- *WCM baseline* = the first year that data are compiled for all measures. Ideally, the WCM baseline year would be the time of wilderness designation. Realistically, the WCM baseline year will likely be the first year that WCM is implemented because wilderness areas generally lack data from the time of designation for most—or even all—measures. The status of wilderness character in **the baseline assessment is the reference point against which change in wilderness character is assessed and evaluated over time.**
- *Measure baseline* = the first year that data are compiled for an individual measure. It is **the reference point for evaluating the trend in a measure over time.** While the measure baseline year will often be the same as the WCM baseline year, it may predate the WCM baseline year if legacy data are used, or it may post-date the WCM baseline year if the data source or data protocol change. The first value reported for a measure from this measure baseline year is called the **measure baseline value.** If new or better data or data sources become available over time, it may be appropriate to adjust the measure baseline value. For example, improved data may result in a recalculation of the miles of system trail from 30 miles to 25 miles, but would not indicate a change on the ground or an improving trend in the measure; instead, 25 miles should become the new measure baseline value. In this situation, the measure baseline value would be reset based on the best available data, and future trends in the measure would be assessed against the time of the new measure baseline value. In all cases, any such change must be clearly documented and rationalized. Before resetting an existing measure baseline value, park staff are encouraged to consult with resource specialists, the appropriate NPS Regional Wilderness Coordinator, and/or the NPS WASO Wilderness Stewardship Division to ensure the appropriateness of this action. Guidance on interpreting the impact of resetting the baseline value of one or more measures on WCM baseline conditions will be developed.

2.2 Defining or selecting measures

Measure selection is described on pages 12-15 in the *Draft Technical Guide*, with a description of desired qualities of measures (useful, simple, and practical), as well as the list of *strongly encouraged measures*. The use of *strongly encouraged measures* is not required, but the following arguments for their use are

provided: (1) pre-established protocols to support easy compilation, (2) “effective WCM can be nuanced, and parks can avoid common pitfalls by using the strongly encouraged measures, which have been vetted and refined by subject matter experts, use the best available science, are endorsed by the NPS WASO Wilderness Stewardship Division, and have been field-tested by parks around the country,” and (3) a “nationally supported protocol has been developed for each of the strongly encouraged measures and use of these measures provides increased national consistency to WCM.”

While these are important considerations, the phrase “effective WCM can be nuanced” can also be used as an argument against automatic adoption of the strongly encouraged measures, particularly for parks with unique characteristics such as GAAR and other Alaska parks. Vetting a measure in one or several parks using “best available science” does not imply it is vetted for any park and the generalization very much depends on similarities among parks. There are also statistical aspects to some of the recommendations that are not generally agreed to be “best available science.” Selecting and justifying context dependent measures should outweigh the value of consistency, ability to use pre-established protocols, and trusting a “vetting” process when extrapolating to new parks.

The “practical” aspect of measure selection states that measures should be chosen that “have pre-existing data from national, regional, or local monitoring programs, or are cost-effective if new data collection is needed.” The reliance on data already collected for other purposes introduces additional challenges through constraining the possible measures, perhaps unintentionally. Choice of measures based on data collected for another purpose should also not be assumed to meet the objectives and goals of WCM, even if they are expected to meet the goals of other monitoring programs. Data adequacy should always be assessed relative to the current goal, as recommended in the Draft Technical Guide.

Conclusions and decisions within the WCM framework rely on the choice of measure *and* the assigned threshold for change which specifies the amount of change that qualifies as meaningful in terms of wilderness character. The process of selecting a measure may be more efficient and productive if the threshold for change is considered simultaneously, as thinking through the *threshold for change* can provide key insights into advantages and disadvantages of potential measures. The next section discusses the *threshold for change* part of the process in more detail, as it is connected to most of the statistical wording and recommendations found in the *Draft Technical Guide*.

2.2.1 Summary and recommendations

Summary WCM does not intend to plan or implement new sampling efforts to collect data for measures; instead, data collected for other monitoring efforts or objectives are to be used to avoid duplication and encourage efficient use of resources. This restricts the possible options for measures for each indicator, which can introduce additional challenges and limitations in terms of aligning measures with specific WCM indicators (i.e., how well does a measure capture what it’s meant to measure?).

Recommendations

- This approach may be necessary for logistical and cost reasons; however, the limitation introduced by the restriction should be explicitly discussed in the WCMPs. Types of limitations include assessing how well the chosen measure actually captures information about the indicator it’s meant to measure and addressing issues related to data collection and scope of inferences (e.g., difficulties in separating changes over space from changes over time).

- As a general strategy, it is recommended that parks first identify what measure they would use *if* resources for data collection were unlimited. That is, rather than starting from the list of available data and measures, start the process with deep consideration of what would be ideal, or at least closer to ideal. Then, work to keep as many of the characteristics of the “ideal” measure as possible when constraining the final measure based on available data and resources. Documenting differences between the ideal measure and chosen measure may help in describing and understanding limitations and challenges relative to wilderness character.
- Also consider other sources of information that might not be considered when referring to “data,” such as professional knowledge of the particular ecosystem by subject matter experts.

2.3 Defining a threshold for meaningful change

A fundamental aspect of the Wilderness Character Monitoring (WCM) plan is the definition of what change in a measure should be taken as meaningful from a practical perspective. The *Draft Technical Guide* (pg 19) introduces *Threshold for change* with the following statement: “This section identifies the **amount of change necessary to qualify as a meaningful change** in the measure” (also see excerpt in Section 2.1 of this report). This sentence does an effective job capturing the goal and should be explicitly kept at the forefront of discussions when attempting to justify a particular threshold. While this recommendation may sound obvious, it is more subtle in practice, particularly when statistical methods and arbitrary criteria are relied on for automatically or implicitly defining the actual threshold.

2.3.1 Challenges and benefits associated with the *threshold for change*

The defining of meaningful change and the choosing of a threshold are integrally linked, by definition. The process of defining a meaningful change can provide a foundation for other parts of the WCM process. That is, the process of explicitly defining a magnitude of change in a measure that *qualifies as meaningful* from a wilderness character perspective creates a practical context, or backdrop, to support future analysis, interpretation, justification of results, and decision making.

The process of defining meaningful change should avoid reliance on default or automatic approaches (such as statistical significance) that tend to move the definition too far from the practical context. Unfortunately, relying on such approaches is common in science and can make one perceive it as more rigorous and/or objective than an approach justified through more qualitative reasoning based on professional knowledge or expertise. The goal of objectivity is understandable, but the process of detaching a definition from its intended meaning or context is a large price to pay in situations such as WCM where results may influence stewardship decisions; it can inadvertently disconnect the underlying objective of tracking wilderness character from the definitions and criteria employed.

Admittingly, the goal of quantifying and monitoring the “character” of wilderness is a challenge, and the difficulties associated with it should be explicitly acknowledged throughout the process. The process may require incorporation of human values and other considerations that are not taken into account if the process relies too heavily on statistical methods and criteria. For the monitoring to be most effective and results to be useful, human judgment will be needed for most, if not all, measures and their associated thresholds for change. The ability to provide convincing justification of choices to people with knowledge of the system should outweigh the value placed on efficiency and ease that comes with choosing default measures and thresholds.

2.3.2 Approaches to defining a *threshold for change*

An example of keeping the measure integrally connected to the threshold and context of the problem is the development of a categorical measure with categories defined by meaningful changes; any shift in category is interpreted as a meaningful change in wilderness character. An example of this is provided within the description of the *any change* threshold option within the *Draft Technical Guide* (also provided in Section 2.3.3.1 in this report). The example is in the context of invasive plants and the categories are based on information about area covered, but also professional judgment. The justification is transparent, with the connection between the measure and the definition of meaningful change clearly defined and open for discussion and debate.

An approach such as described in the previous paragraph likely requires more work in the WCM planning phase than would be required for more default approaches discussed below, but should require less work and justification later in the process. For example, the invasive plants example likely took a fair amount of effort to get experts to agree upon and justify the categories, but the assessment in the end will be simple and relies on few, if any, additional assumptions.

Using an automatic statistical analysis falls at the other end of the spectrum. Stating “we will rely on regression analysis and an alpha level of 0.05” requires little work up front, but a good deal of work at the end to assess assumptions and attempt to translate results into meaningful interpretations and decisions in the context of wilderness character – if this last important step is actually undertaken to back up conclusions.

The statistical modeling and a default p-value cutoff approach to define meaningful change (associated with a null hypothesis significance test and a desired Type I error rate) disconnects the *threshold for change* from the practical backdrop and over-relies on using the precision of an estimate as the criteria as opposed to also the magnitude, as well as inserts potentially strong assumptions through an underlying model (such a linear change in the mean). Language in the *Draft Technical Guide* favors this approach for its supposed “rigor” over less statistical approaches; this implication is challenged in this review and discussed in greater detail throughout the report. The approach may be justifiable for some measures and contexts, but should require justification as to why it is chosen as the preferred approach, as opposed to selecting it as a default with little or no expectation of justifying the decision.

A helpful litmus test is to (hypothetically) go through the process of justifying to a stakeholder, or interested party with little or no statistical background, why a result based on a statistical model and hypothesis test should be taken as the basis of a conclusion regarding meaningful change in wilderness character. This exercise should point to the underlying problems of the approach in a qualitative context such as WCM. Unfortunately, a norm in some areas of scientific practice has become the lack of expectation of justification when statistical methods are used, which adds to their appeal (even if subconsciously) while perpetuating deep underlying problems. Justifications that are given are usually superficial, such as blanket statements of “increased rigor” or “increased objectivity.” The practical implications of relying on an approach that is not well justified in the context of the problem are potential mistakes in stewardship decisions, or least conclusions and decisions made without strong justification or sound reasoning.

2.3.3 Three options in national guidance

The description of the protocol for setting thresholds for meaningful change currently described in the *Draft Technical Guide* and *Draft GAAR WCMP* (provided in Section 2.1 of this report for reference) specifies three *standard categories* to be used for thresholds: (1) Any change, (2) Percent change, and (3) Statistically significant change based on regression. Each of these options is discussed and briefly reviewed within this section.

2.3.3.1 The “any change” option

The first standard category used for thresholds is *Any Change*. This option is the most straightforward and therefore easiest to convey to stakeholders and for others to critically evaluate. Its usefulness is integrally connected to the associated measure.

The *Draft Technical Guide* states that the *Any Change* option “is typically used for measures where change over time is unlikely and where there is high certainty about changes in the data.” While this statement provides a useful starting point, it can be expanded to better align with the idea of “specifying the amount of change that qualifies as meaningful.” It is possible to have measures “where change over time is unlikely” but where “any change” would still *not* qualify as meaningful in terms of changing the underlying wilderness character. Likewise, degree of uncertainty about changes in the data also does not directly inform what would qualify as meaningful; it is possible to have a measure with little uncertainty (“high certainty”) in changes and still not agree that “any change” should qualify as meaningful. For example, a count of non-native plant species in a small area frequented by humans may be considered to have little error and a change over time might be unlikely, but a change by one species may not be considered meaningful change in terms of wilderness character. Therefore, how likely a measure is to change over time and the degree of uncertainty in the data are important to consider, but they still leave the crucial part of a threshold for change - the “amount of change that qualifies as meaningful.”

Interestingly, an *Index of Invasive Plant Species* (not the *strongly encouraged measure* described elsewhere) is used as an example in the *Threshold for Change – Any Change* section (see excerpt in Section 2.1). The index described is constructed using categories that were created based explicitly on definitions of what qualifies as meaningful change for the indicator. Because the categories are directly defined by meaningful changes, a movement in category over time is “by definition” a meaningful change. In this way, the threshold for change is built into the measure itself, rather than viewed as a separate part of the process overlaid on a change in a quantitative measure to judge if it is meaningful. This idea is discussed further in Section 3.5.1 and is an important application of the *any change* threshold option. The idea does not apply to any measure defined by categories, as the categories may not be directly tied to a definition of meaningful change and separate work for the threshold for change is still needed (*any change* may not necessarily be a reasonable choice). For example, a shift by one category may not correspond to meaningful change, but maybe a shift by two categories may. It is also possible that the same shift does not apply across all the categories, though it may be possible to combine and/or split categories to align category shifts with meaningful changes and justify the *any change* option.

In summary, the *Any Change* threshold can be an effective, transparent, and easy to evaluate threshold. Things to check are : (1) Do categorical measures have categories defined such that any shift in category

is considered meaningful? (2) If an Index or other quantitative measure has no upper bound, is *any change* appropriate within any range of values the measure can take on? (e.g., for the *Index of Invasive Plant Species* described as the strongly encouraged measure, would a change from 2 to 3 be considered the same as a change from 15 to 16 in terms of meaningful change?), and (3) Considering temporal variability and data adequacy are important, but are not the only factors that should lead to the choice of *Any Change* – an assessment of magnitude of change in the measure qualifying as meaningful should also be included.

2.3.3.2 The “percent change” option

The second standard option for specifying a *threshold for change* is *Percent Change* using either 3%, 5%, or 10%. The document implies that if a decision is made not to use *Any Change*, then *Percent Change* is the next logical consideration for any measure. Several concerns with this approach are identified and briefly discussed in this section.

The *Draft Technical Guide* recommends the use of the *percent change* option for “measures that are less sensitive to change, that show variation from year to year, or that have medium data adequacy. Larger percentages indicate a higher likelihood of annual variation or a lower expected data adequacy.” Similar to the advice given for the *any change* option, the stated considerations of yearly variability and data adequacy are important, but they should be considered layers on top of the foundation of specifying a magnitude of what qualifies as meaningful change in the underlying measure. It can be helpful to think about what changes in the “true” measure would be meaningful - before considering measurement and sampling error. Statistical reasoning can then be helpful to considering detection of that meaningful change relative to potential sources of error and variability in measurements over time and space.

The progression from *any change* to *percent change* is overly restrictive as it leaves out an important alternative, the option to define meaningful change in terms of an additive increase or decrease. Presenting *percent change* as the only option between *any change* and *regression and statistical significance* implies that a multiplicative change is the only other option. For example, it may be that it makes most sense to define a meaningful change by an increase of 5 points (additive), rather than an increase of 5% (multiplicative). This decision is often dependent on the properties of the underlying measure and the range of values that may occur for a given wilderness area. The distinction between defining meaningful change in terms of additive or multiplicative is an important concept in defining meaningful change and choosing statistical models, if desired.

The restriction to options of 3%, 5%, or 10% is hard to justify. There is no reason that all measures should fall into one of these choices for defining meaningful change. Instead, the choice of percent change should be measure and context specific. Rather than setting up the process to choose one of the options with little expectation for additional justification, any choice for percent change should be encouraged with justification. As discussed in further detail in other sections, recommending default choices may tend to encourage lack of rigorous justification for the ultimate choice and disconnect actual meaningful change in a measure from decisions made in the WCM process.

An important consideration is the starting value that will be used with the percent change approach. For example, in GAAR the Index of Invasive Plants or count of number of non-native species has a baseline measure value of 0. This may be the case for other measures in Alaska as well given the pristine nature of the parks. A starting value of zero automatically rules out percent change as a reasonable approach – given that a 5% change in zero is zero. In such a case, *any change* or *additive*

change would be more appropriate. The starting-value problem does not simply disappear when measure values are greater than zero. For example, measures that are low counts will also have problems with the percent change options. Suppose the starting value is 1 (e.g., there is one non-native species known to be in GAAR). Then, a 5% change is $1 + (1 \times 0.05) = 1.05$ or $1 - (1 \times 0.05) = 0.95$. For a starting value of 2, the thresholds are 1.9 and 2.1, and for 3 they are 2.85 and 3.15, etc. So, if comparing a count to these non-integer thresholds, these criteria are no different from “any change” or an “additive change of 1.0 species.” The threshold does not change from *any change* or an additive change of 1.0 species until the starting value is greater than 20. It is potentially misleading, or at least not fully transparent, to describe it in the “percent change” context and is more straightforward to a broad audience to simply define the threshold in terms of the *any change* or an *additive change*.

It is also important to specify whether the percent change is defined relative to a single data value for the measure or a summary statistic of multiple data measures. This will depend on how the baseline measure value is obtained, as well as how future values are obtained and summarized over time. Care should be taken to make sure the definition of meaningful change takes this into account (e.g., meaningful change should not be defined on the scale of single data values and then applied to averages). For example, the above example assumed a new single count of total species would be compared to the baseline measure value (also a single count), rather than an average count over multiple years.

The web applet described in Appendix A.2 provides a tool for visualizing user-specified additive and multiplicative changes per year (or 5 years) based on a given baseline measure value. The visualization can help provide context for discussion about what changes may be realistic and what changes should qualify as meaningful.

It is not always clear in the wording within *KIW2*, the *Draft Technical Guide*, and the *Draft GAAR WCMP* whether the *any change* or *percent change* threshold are meant to always apply back to the baseline measure value (as implied in the definition of *Measure baseline* on page 18 of the *Draft GAAR WCMP* quoted in Section 2.1 of this review) or if future values will be compared against the most recent or some aggregation of all past measures (as is the case if regression is used). This decision has practical consequences for the process and for results and their interpretation.

2.3.3.3 The “statistically significant change based on regression” option

The third standard option is “statistically significant change based on regression.” In this section, several issues are identified, with Sections 2.4 and 2.5 providing more detail. Recommendations in this review, from my expertise as a statistician, may differ from norms in many scientific disciplines in terms of the automatic use of statistical methods and default thresholds to support conclusions and decision-making.

Some of the concerns raised in this and subsequent sections have potentially serious implications for conclusions coming out of the WCM process; conclusions are expected to be sensitive to approach because estimation of change and threshold depend on them. A very common misunderstanding of statistical analysis is the belief that “statistical significance” indicates practical relevance or importance. Statisticians have been warning about this well entrenched belief for decades, but it is difficult to reverse practices that become engrained as part of scientific culture. The use of statistical significance in the context of WCM as a default way to “define meaningful change” is a classic misuse of statistical inference.

A paper by Greenland et al. (2016) provides an accessible and succinct overview of many of the issues related to some of the major myths associated with p-values and statistical significance, much of which is relevant to language used and recommendations given in the *Draft Technical Guide*. The Greenland et al. (2016) paper is part of a large and varied body of research and publications cautioning against misuse of statistical hypothesis tests and thresholds, such as the other papers included in the same supplementary issue of *The American Statistician* and the American Statistical Association's *Statement on P-values*.

The *Draft Technical Guide* states that “For more complex measures, some form of statistical analysis may be needed to determine the change threshold.” This statement does not provide adequate justification for choosing to rely on statistical models and tests; it also does not imply that regression and statistical significance are appropriate. First, it is not clear what constitutes “complex” in this situation. Is it referring to properties of the measure (such as substantial variation, data coming from a complex design, substantial measurement error, etc.)? Or, is it referring to challenges in justifying what amount of change qualifies as meaningful? Regardless of the intended meaning, simply having a more complex measure does not imply that statistical analysis and significance tests are automatically appropriate or “needed”.

In the context of WCM, the statement that statistical analysis can “determine the change threshold” implies that statistical analysis and criteria for “statistical significance” are capable of “determining” what constitutes a meaningful change in a measure relative to wilderness character. Use of statistical modeling, a null hypothesis significance test, and specification of an alpha level may be a convenient way to define a threshold, but there is no theoretical justification as to why that threshold should correspond to “the amount of change necessary to qualify as a meaningful change in the measure.” Using it in this way encourages bypassing of the often challenging task of justifying a definition of meaningful change.

As further discussed in Section 2.5, statistical reasoning can help judge the potential for meaningful change relative to background variability, but that requires a separate part of the process to justify what magnitude of change in the measure should qualify as meaningful. That is, the justification of what qualifies as meaningful change from a practical perspective in the context of the problem is a step that should accompany (really precede) any statistical analysis, not a step that should be given over to statistical methods. To get the most out of a statistical analysis used to support practical decisions, the analysis should be supported by a well justified backdrop of contextual relevance – a backdrop where magnitudes of change considered meaningful are defined *before* analysis. The WCM process is a perfect opportunity for this given the explicit definition of the threshold and emphasis on creating a WCMP. Side stepping this crucial part of the process to rely on default statistical criteria with no explicit connection to the context is unfortunate and leads to less effective and trustworthy conclusions from the WCM process.

2.3.4 Summary and recommendations

Summary A *threshold for change* chosen according to the options presented in the *Draft Technical Guide* (pages 19-22) may fail to represent meaningful change in an indicator of wilderness character. Conclusions regarding trend depend on choice of threshold, making specification of the *threshold for change* a crucial part of the WCM process.

Recommendations:

- The step of defining meaningful change should rely on professional expertise and context for justification and precede statistical analysis. Statistical tests and p-value cutoffs are not designed to answer the question of what amount should be judged as “meaningful.”
- Rigor can be improved by setting an expectation for justifying the *threshold for change* to a broad audience, using the scale of the chosen measure and aligning with the definition of wilderness character.

Summary The *Draft Technical Guide* emphasizes the importance of “measure selection,” but downplays the importance of assigning associated thresholds.

Recommendations

- Defining a *threshold for change* should be included explicitly in Step 2 of the WCM process (pages 16-17 *Draft Technical Guide*).
- The reasons for choosing a particular *threshold for change* should be transparent and explicitly justified relative to the measure and the ultimate decision about trend, rather than presented as if it is more of a technical aspect of the analysis process.
- Usefulness of a conclusion regarding trend in an indicator of wilderness character depends on the interplay between the measure and the threshold for change. Therefore, choice of measure and *threshold for change* should be considered simultaneously.

Summary The three options for *threshold for change* presented in the *Draft Technical Guide* (“any change,” “percent change,” or “Statistically significant change based on regression”) are overly restrictive and do not encourage context specific choices.

Recommendations

- The *Draft Technical Guide* describes flexibility in choice of measure afforded to each NPS wilderness area; that flexibility should also apply to defining *thresholds for change* given that conclusions will be sensitive to this part of the process. In general, terms such as “strongly encouraged measures” and “three options for defining threshold for change” may discourage deviation from the recommendations even when justified.
- Consider including other options in the *Draft Technical Guide*, such as “additive change,” as well as encouraging creativity in developing categorical measures and associated thresholds simultaneously.
- Basing a *threshold for change* on “statistically significant change based on regression” is not recommended for use as a default.
- Build in an expectation to spend more effort in the planning phase to develop park-specific measures *and* thresholds, with the goal of building a strong foundation leading to well-supported conclusions with less need for justification later in the process.

Summary Using “statistically significant change based on regression” as the method for setting a *threshold for change* has severe limitations in its usefulness and defensibility relative to the goals of WCM. The preference given to this option seems to rely on an unstated assumption that if at all possible, regression and statistical significance should be used because it provides the most rigorous approach available. This assumption is not justified or defensible.

Recommendations

- Remove statistical hypothesis testing as a default approach for defining a *threshold for change*. In general, use of statistical methods does not automatically increase rigor of the WCM approach, as implied in the national documents. Relying on statistical methods and commonly used cutoffs is often viewed as more scientifically rigorous than relying on approaches perceived

as less complex or sophisticated; however, “rigor” relies on how well the model is aligned with the practical questions and goals, and the extent to which model assumptions are justified.

- Revise wording that implies the use of statistical methodology renders an approach more rigorous and/or useful than other options. Statistical approaches often carry a perception of “objectivity” and lack of “bias” that is rarely questioned. It can be attractive to rely on such an approach, particularly when it makes the process more automatic and easier to implement -- not only in carrying out the calculations, but because it often comes with fewer expectations in terms of having to justify one’s approach.
- Revise specific wording that is not well justified statistically. For example, complexity of a measure and availability of at least five measure values are not well justified reasons for use of linear regression, and “statistical rigor” is not gained simply by using a lower alpha-level.
- The process for specifying a *threshold for change* should stay aligned with the goal of defining and justifying what amount of change in a measure represents meaningful change in wilderness character; a statistical model and test do not automatically provide that information.
- Instead of placing preference on a statistical hypothesis test, it is recommended that preference be placed on approaches explicitly justified and able to be critiqued by stakeholders (see next topic).

2.4 Considering the role of statistical analysis

After the process of defining meaningful change, then statistical analysis can be a valuable tool to assist with inference in terms of taking into account variability in the measure over time to estimate the mean and quantify some sources of uncertainty combined (e.g., measurement error and natural variability over time and space). Such an approach first relies on connection the definition of meaningful change to a parameter in the model that will be estimated (e.g., the slope parameter in simple linear regression). The estimate and confidence intervals (or compatibility intervals or credible intervals) can be used to evaluate measured changes or estimated trends to the meaningful change backdrop created pre-analysis. This allows for simultaneously considering background variability in the measure (as captured by statistical estimation) and the amount of change that qualifies as meaningful. Statistical analysis does provide a way to help assess magnitude of an estimate relative to background variability and can be very useful if interpretation honors the definition of meaningful change specified beforehand and assumptions made in the statistical modeling are considered. It is not, however, meant as a way to define what qualifies as meaningful change for a practical problem. See Section 2.5.4.4 for an example of this process created as part of a review of the Landres et al. (2009) Appendix B which is the reference to support the use of statistical modeling and hypothesis testing in KIW2 and the *Draft Technical Guide*.

The *Draft Technical Guide* states: “If a more involved statistical analysis is required to assess trend, the assistance of a statistician may be necessary, and is encouraged.” It is not clear what would lead to it being “required,” but it is important to acknowledge that a more involved statistical analysis is always a choice that brings with it the responsibility to consider the many additional assumptions that should be assessed. A more involved statistical analysis should not take the place of the up-front work needed to justify what should qualify as meaningful change in a measure relative to wilderness character. It also requires work to connect the quantity, or measure, of interest to a parameter defined within a statistical model and to thoroughly check that assumptions are reasonable.

The role of statistical analysis should also be considered in the context of transparency of methods, assumptions, and decisions to the broad collection of stakeholders with varying backgrounds. When an

automatic procedure is one that many stakeholders do not have the expertise to critically evaluate (e.g., linear regression and statistical significance), justification is allowed to be superficial and the approach may be given more trust than it deserves, particularly when it is described as the most “objective” and/or “rigorous” approach. Instead, rigor can be improved by setting an expectation for justifying the threshold of change to a broad audience using the scale and context of the chosen measure. This does not mean that statistical analysis is never an option, only that it is considered on a case-by-case basis and thoroughly explained and justified, rather than chosen automatically.

2.5 Use of linear regression

A substantial part of this review focuses on the language and recommendations in national guidance documents related to the use of linear regression and statistical significance for estimating change and judging whether a change is “meaningful”. This part of the WCM process deserves for multiple reasons: it applies across parks and measures; it is a relatively easy part of the process to revise; and it serves to promote awareness and discussion about the role(s) of common statistical methods in science and monitoring.

2.5.1 Regression models for assessing trend

For WCM, “statistically significant change” is described in the context of a simple linear regression (SLR) model is appropriate and testing the null hypothesis that the slope parameter is equal to zero. The SLR model is defined using year as the explanatory variable (x-axis) and the chosen measure as the response variable (y-axis variable). This is a common framing of statistical modeling of change over time and testing of trends, although it varies in appropriateness depending on how reasonable the model is and how the hypothesis test is interpreted, along with what other information is used. The comments in this review also apply to multiple linear regression, where other covariates or predictors are included along with year, though there are then additional assumptions to assess.

The *Draft Technical Guide* states: “the regression threshold applies to certain measures once they have accumulated a sufficient amount of data and is used because it provides statistical rigor in the long-term analysis of trends in the measures.” First, it is important to reiterate here that a linear regression model and associated null significance hypothesis test does not simply buy statistical rigor, even for long-term analysis. The rigor associated with the approach depends on justification for its use, the reasonableness of the assumptions underlying the model and the test, and how the results are used and interpreted relative to the goals of the WCM process.

Using linear regression is not necessarily more “robust” than other approaches. The term “robust” is often used in both technical and non-technical ways and its intended meaning should be justified. For example, the following sentence from the *Draft Technical Guide* and *Draft GAAR WCMP* is not technically true and depends on the definition of “trend” and the definition of “robust:” “In a given year that trend is reported, measures that have at least five data points, simple linear regression should be used to determine trend. For measures where the five-year minimum monitoring interval yields only one measure value, the “trend” in a measure is determined by comparing the most recent measure value (most commonly a three-year rolling average) with the earliest available baseline data for that measure. While this trend is not as robust as the linear regression analysis, it will give managers a preliminary sense of measure trend until such time that five data points are available.”

A distinction should be made between relying on a regression model to estimate the change in the mean of a measure over time vs. relying on regression to “determine the change threshold.” It is possible to use a regression model to assist with estimating trend, while not basing the definition of meaningful change on a hypothesis test associated with the slope parameter and a specified alpha-level (discussed further in the next section). Recognizing the layers involved is important in understanding the potential problems with the approach. In this section, the simple linear regression model is discussed. The hypothesis test and “statistical significance” pieces are addressed in Section 2.6.

Adoption of the linear regression model (as recommended in several places in the guidance) should require careful consideration and evaluation of the assumptions underlying the model, even before considering the recommended hypothesis test. Having “a more complex measure” does not imply the regression model will be appropriate in its underlying assumptions. WCM protocol is to have regression implemented automatically within the WCMD, meaning assumptions will clearly not be adequately assessed and any conclusions should be questioned.

2.5.2 Assumptions of the linear regression model

As for any statistical model, the assumptions are never “met” – but are always violated to different degrees. Many statistical models are robust to moderate departures from some of their assumptions, meaning they are still “valid” according to statistical criteria even under noticeable departures from the assumptions. Robustness to violations depends on the assumption. Often, assumptions are checked using plots of residuals or even using other statistical tests with their own assumptions (that are left unchecked - even when potentially less robust to departures than the original model and assumption). The end result in practice of these exercises is often a “met” vs. “not met” declaration. This approach is convenient, but overly simple and ignores existing nuances that can greatly impact conclusions.

There is also an added worry about assumptions within the WCM process because of how the approach is to be carried out technically. The intent is to give the WCMD (a computer program) the information that it should use linear regression and a p-value cutoff before data are entered. Then, when measure values are input into the system, the estimation and thresholding will be done automatically behind the scenes. This creates a dangerous system where the crucial step of checking assumptions is never undertaken. It is not clear where the human should enter this part of the process to undertake this important step. As discussed further in Section 2.9, this is a undesirable implication of the desired automation of the process.

Beyond the usual assumptions often checked by data analysts (e.g., normality), there are more fundamental decisions that should also be considered, such as the decision to base the analysis and subsequent decisions on statistical models and criteria to begin with. The WCM process presents an interesting insight into the perceived role of linear regression and hypothesis tests through the progressions provided in three options for specifying a *threshold for change*: any change, percent change, then linear regressions and “statistical significance.” It is useful to step back to consider the logic behind this progression and what changes are made in the underlying assumptions when moving from percent change to regression, for example, as is suggested for many measures (see Section 2.5.5 for more discussion on this).

The following properties and assumptions should be considered or assessed before relying on the linear regression model for inference or decision making:

- *Shifting the approach*: Moving to use of a linear regression model to define “trend” in terms of an estimated slope moves away from a direct comparison to a baseline measure value as described as a goal in multiple places of the WCM guidance. This may be a desired change in approach, but the transition away from comparing to the baseline value should be acknowledged and justified for each measure. The existence of 5 values is not considered adequate justification.
- *Model for the mean*: The regression model assumes interest lies in estimating how the *mean* of the measure changes over time. This is fundamentally a different approach from reliance on strategies such as Any Change or Percent Change with reference to the baseline measure value. That is, assessing change based on raw measure values is different than basing monitoring on an average over multiple values, or an underlying mean over time. The choice should depend on the measure and be explicitly acknowledged and justified. If the approach is going to be changed mid-monitoring (such as switching from percent change to regression after 5 values), the reason for the shift should be explained. In general, throughout the guidance documents there is a lack of clarity and consistency regarding what measure values or summaries will be used to estimate change.
- *Linearity*: The SLR model assumes the mean of the measure changes linearly over time for the duration of the monitoring (each yearly increase is associated with the same additive increase in the *mean* of the measure). Linear regression using the measure on the log scale provides a model assuming the same multiplicative increase each year, although it is not clear if this is supposed to also be an option for WCM. The assumption of linearity should be assessed using a plot of the observed measure values, as well as an *a priori* consideration of the problem and whether such a simple relationship will be sufficient and adequately capture change in measure (and ultimately indicator) over time.
- *Normally distributed residuals*: For a given year, the hypothetical collection of values of the measure are approximately *normally distributed* around the fitted regression line. This is conceptually difficult to consider because a single year does not have multiple values; in practice, the collection of residuals (difference between an observed value and the regression line) are assessed together. This assumption is difficult to assess when there are few observations, as expected for WCM. However, severe violations will likely be easy to identify and many can be anticipated before data are collected just based on properties of the measures. For example, if the measure is a small count, the assumption of normality is expected to be seriously violated (e.g., a count of non-native species).
- *Variance of residuals is constant*: An assumption related to that of normality is that the variance of the assumed normal distribution has the same variance across all years. In other words, the variance of observations around the fitted regression line is “constant” over time (i.e., constant variance). Inferences can be sensitive to violations of this assumption, affecting the width of confidence intervals and magnitude of p-values. This assumption can be assessed graphically and through consideration of the properties of the measure in question (e.g., counts of species are not expected to exhibit constant variance over time if the mean of the counts is increasing over time).
- *Independence of residuals*: The regression model assumes the observations of the measure are *independent* after accounting for the linear change in the mean over time (independence among residuals). This is usually the most challenging assumption to assess because there is no single graph or test. In practice, there are always reasons to suspect lack of independence and it is mainly a thought exercise to identify them. For example, an obvious one in the context of estimating trend over time is temporal dependence – the idea that observations closer in time are more similar than observation further apart in time. In terms of residuals, this means that residuals for observations close together in time may tend to be positive or negative together (i.e., falling on the same side of the fitted regression line). Another common source of dependence is due to closeness in space,

rather than time. If sampling region changes over time, observations from the same or more similar regions are expected to be more similar than independent observations may be. (A separate but related worry is the confounding of changes over space with changes over time if sampling regions change over the course of the monitoring). Estimates of statistical uncertainty based on linear regression are sensitive to violations of the independence assumption; positive autocorrelation in the residuals can manifest as standard errors that are smaller than they would be under independence (meaning confidence intervals are misleadingly narrow and p-values are misleadingly small). Degree of autocorrelation in the residuals can be estimated and corrected for by using more sophisticated models, but it is very difficult to estimate and correct for when a time series is short (less than about 50 observations), as will be the case for WCM for some time to come. Again, this does not imply regression should not be used, but it does suggest results should be interpreted with caution and p-values should not be compared automatically to cut-offs to make black-and-white decisions regarding whether meaningful change occurred.

- Many examples can be created where the comparison of an individual year's measure value against the baseline measure value would indicate an upward or downward trend in the measure, but reliance on the estimated slope of linear regression would lead to reporting a stable trend. The opposite inconsistency is also possible. This is not to say one approach is always preferred, but it is important to recognize potential inconsistencies and understand the reasons for them in order to better justify the approach taken.

Connecting a threshold for change to a null hypothesis significance test for the slope inherently includes additional assumptions and considerations. These are discussed in Section 2.5.4.

2.5.3 Summary and Recommendation

Summary The decision to rely on a simple linear regression model should not be taken lightly. A crucial part of the process is carefully assessing the assumptions before using the model. Without this, it is difficult to assess to what degree the end results and conclusions are influenced by the information contained in the model assumptions vs. the information contained in the data.

Recommendations:

- If regression is going to be implemented automatically within the WCMD, add an explicit step to the instructions that requires the user to assess assumptions and provide written justification for them before regression is used.
- Do not present “regression and statistical significance” as a method that is appropriate for any measure with 5 or more values.

2.5.4 Threshold for change and the use of regression

A key aspect of the use of regression to estimate “trend” is the availability of a statistical hypothesis test with a null hypothesis of the slope equal to zero (often interpreted as “no trend”). The third *threshold for change* option in the *Draft Technical Guide* is based solely on using the p-value from the hypothesis test and comparing it to a pre-specified cut-off described as an “alpha level.” Therefore, the threshold part of the process is completely contained within the statistical modeling and testing. This section discusses some of the issues and concerns with this approach.

2.5.4.1 The threshold changes over time

The WCM planning process relies on clearly specifying a baseline measure value and a threshold for change for categorizing trends, with the expectation that if either is changed the reasons will be documented and justified. When the criteria are stated as comparing the p-value to 0.1, for example, it can appear that the threshold stays the same over time, as intended. However, one of the potential problems with using “statistical significance” is that the definition of meaningful change associated with a p-value-based threshold will change over time based on the number of values used in the analysis, the variability in the observed values, and how the values change over time. P-values are sample size and model-dependent.

For most measures, changing the definition of what qualifies as meaningful change over time does not make sense, and therefore this should be justified. Additionally, the change is hidden behind the methods and is not easy to describe or critically evaluate, further contributing to lack of transparency in the process when using the regression approach.

2.5.4.2 Translating a p-value threshold into a slope threshold

As discussed previously, it is challenging to translate a p-value-based threshold back into an amount of change in the measure. Using this approach disconnects the “qualifies as meaningful change” from the context of the measure and its connection to wilderness character. Additionally, the trustworthiness of the p-value is unknown given the unknown extent of the violations of assumptions (as discussed in Section 2.5.2). This approach is not recommended as a default for the WCM process.

Given common reliance on this approach in practice, a resistance to this recommendation is expected. It may be helpful, as an exercise in understanding, to translate between the p-value and the magnitude of the slope parameter in the regression model. The slope can be more readily tied to the context of the measure, although it still relies on the assumption that change in the mean of the measure will be the same within any year. This translation can be done mathematically for simple linear regression *under additional assumptions* and assuming all assumptions of the simple linear regression model are perfectly met (which we know is not the case in practice). Therefore, results of the translation should be interpreted with caution and with an attitude of learning about the method, not obtaining “an answer.”

To facilitate understanding of the connection between these thresholds, a web applet is provided in Appendix A.4 for stakeholders to investigate translations between p-values and simple linear regression slope parameters under sets of assumptions. To carry out the calculation, an additional assumption is needed beyond that described in the WCM documentation -- statistical power (desired Type II error rate) or desired width of a confidence interval must also be specified. Sensitivity of the results to these assumptions should always be acknowledged. The goal of the applet is *not* to facilitate continued use of p-value cutoffs as thresholds, but instead to provide an understanding of limitations of the approach and to demonstrate the disconnect between a p-value threshold and defining meaningful change within a practical context. The hope is that increased awareness of the regression and statistical significance black box may encourage consideration of approaches that are more justified and connected to the WCM goals.

2.5.4.3 Alternatives to p-values if using a regression model

If the simple linear regression model is chosen to describe and estimate trend over time, it is recommended that estimates and confidence (or compatibility or credible) intervals for the estimated

slope parameter be used to assess trend relative to practical objectives, rather than relying on a p-value. This is a less automatic method and could not simply be carried out within the WCMD. However, it will result in more useful and better justified conclusions. It does require defining the threshold for change in terms of the slope of the regression line (mean change in the measure per year) and assumes the regression assumptions are reasonable.

As illustrated further with an example in the next section, estimates and statistical intervals can be compared to a backdrop of meaningful change in the measure (or parameter of the statistical model) defined before analysis in the broader context of the problem. The goal is to assess the magnitude of the change relative to sources of uncertainty captured in the statistical analysis, and what is considered meaningful. This can be accomplished while taking into account the complication that there is rarely a hard threshold separating non-meaningful changes from meaningful changes, even though adding a decision-making layer to the analysis often forces a single number definition for the threshold.

2.5.4.4 A closer look at Landres et al. 2009 and connecting regression to meaningful change

The reference provided in the *Draft Technical Guide* as justification for the regression and statistical significance approach is Appendix B of Landres et al. (2009); this is discussed further in Section 2.6.3.1. The example provided Landres et al. (2009) provides an opportunity to work through some of the concepts and suggestions provided in this review, while at the same time better understanding how much the document should be relied upon for justification of methods within WCM. Their illustration of the approach is based on air pollution data from Acadia National Park with “Example 1” using the actually observed data and “Example 2” using an artificial data set created by modifying the observed data (so that the associated p-value ends up being just over the 0.1 cutoff). There is not a lot of discussion around the examples, other than the following:

By examining plots of the two data sets, it appears that a downward trend occurs in visibility for both the true data (example 1) and the modified data (example 2), although more noise occurs in the modified data. Nevertheless, performing the hypothesis test with a 90-percent confidence level (alpha equals 0.1) yields the detection of a statistically significant trend for the true data (example 1) but the failure to detect a statistically significant trend for the modified data (example 2).

It is assumed from the presentation of the example and its use to (presumably) demonstrate the benefit of using an alpha level of 0.1., that the above description is meant to describe desired behavior of the approach – that is, that the approach (SLR model plus hypothesis test) is doing what they want or expect it to do in the context of WCM. Unfortunately, the situation is not that simple in the context of the goals of WCM.

Figure B.1 in Landres et al. (2009) presents scatterplots of the two data sets, but on two different scales that make it appear as if the estimated slopes are about equal and all that is changed is the “noise” around the regression line. Their plot is provided in Figure 1 below. Figure 2 provides a more straightforward plot of the data using the same scale for the y-axes and including both sets of data on the same plot, along with the estimated regression lines. A comparison of Figure 1 and Figure 2 reveals that the story is not as simple as portrayed in the above text or implied by their conclusions and recommendations.

Figure B.1.—True (example 1) and modified (example 2) air pollution data from Acadia National Park plotted to show graphical trend in the data.

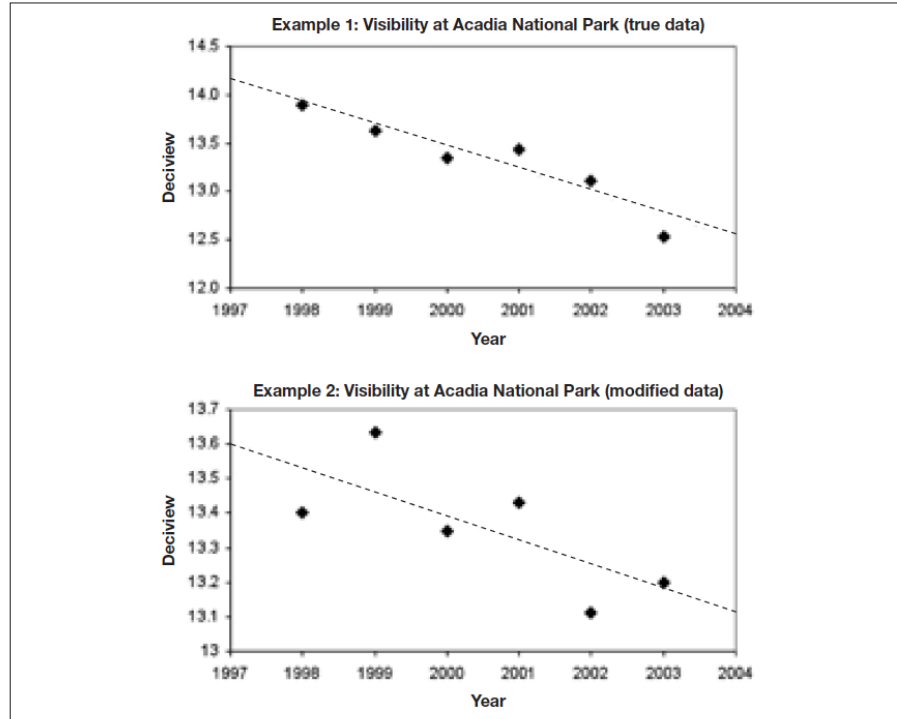


Figure 1 Figure B.1 from Landres et al. (2009) Appendix B.



Figure 2 Re-creation of the scatterplots displayed in Figure B.1 from Landres et al. (2009) so that data are shown on the same y-scale and the difference between the observed data (blue) and modified data (pink) is more explicit (see pink vs. blue points in 1998 and 2003). The claim that the modified data have more “noise” than the observed data is not justified.

First, the observed data are only modified for the first year (1998) and the last year (2003); estimated slopes are most sensitive to changes in these years (points furthest from the middle year have large leverage). The 1998 value is made smaller by about 0.5 and the 2003 value is made larger by 0.7 – both changes contribute to moving the estimated slope closer to zero as compared to the observed data. The middle four values stay exactly the same, as seen in Figure 2 and Table B.1 in Landres et al. (2009).

Second, the modified data do not have more “noise” as claimed by Landres et al. (2009) and supposedly displayed in their Figure B.1. As can be observed in their Table B.2, the estimated variance based on the distance of points from the regression line is actually smaller in the modified data than the observed data. The smaller average residual (vertical distance between each observation and the regression line) for the modified data is readily observed in Figure 2.

While likely unintentional, describing the modifications to the data as only a change in “noise,” while ignoring a change in the magnitude of the slope is misleading, not to mention the plotting on different scales in a way that supports the claim. This is not an example that clearly separates change in background noise from change in slope (though it is possible to use computer simulation to carry out such an investigation over many realizations of data under different assumed models).

The results regarding “statistically significant trends” are based on the simple linear regression model (see Section 2.5.2 for discussion of assumptions) and the null hypothesis of a slope equal to zero. In this case, the SLR model assumptions do not appear severely violated and the model appears to be a reasonable choice, if it is desired to define trend as the change in mean per year. The hypothesis test is set up to conclude either there is a “statistically significant” trend (if the p-value is less than 0.1 or equivalently if the 90% confidence interval excludes zero) or there is a “not statistically significant” trend. This approach is simple and easy and requires little work after fitting the model, as well as little work up front to justify what slope might qualify as meaningful in the WCM context.

The approach of dichotomizing results is set up to ignore all information contained in the confidence interval *except* whether it contains 0. This is an unfortunate waste of valuable information and the process contains no connection to what experts may believe is a meaningful change in deciviews in a year, in terms of a goal for monitoring (like wilderness character). For example, can a human notice a difference of 0.07 deciviews? What about 0.24? How to judge what is meaningful depends inherently on value judgements (view for humans, health of humans, health of animals, health of plants, etc.).

The conclusions provided in Table B.2 of Landres et al. give the following two conclusions: (Example 1) “Reject null hypothesis and conclude a significant trend exists” and (Example 2) “Fail to reject null hypothesis and conclude no significant trend exists.” In Example 2, the two-sided p-value was 0.1035, just barely above the 0.1000 cutoff and the 90% confidence interval was (-0.143, 0.001) – just barely including zero. This was not mentioned in Appendix B, though this example of barely crossing a threshold for observation data is often given as a case against use of cutoffs. Is there really a practical difference between a p-value of 0.099 and 0.110? In this overly simplified approach, the two p-values result in different conclusions regarding trend in this aspect of wilderness character.

Additionally, the wording comes very close to the common misinterpretation of “accepting the null” - while subtle, saying “no significant trend exists” will likely be interpreted by many as “there is evidence for no meaningful trend.” At the very least, any decision making relying on the results will assume “no meaningful change” is the case. The “large” p-value does not provide evidence for concluding “no trend” and whether adding the “significant” word to the sentence changes that interpretation is

debatable. Instead, the “large” p-value simply suggests that given the approach and all its assumptions, the possibility of a zero slope cannot be rejected, or that the data are compatible with a slope of zero (as well as many other slopes that might be practically meaningful). See Greenland et al. (2016) for more on interpretations.

The example provides a good opportunity to go through a process of specifying meaningful change as “amount of change in the mean per year that qualifies as meaningful,” thus tying it to the slope parameter, and demonstrating more holistic and context-dependent use of information contained in the confidence (or compatibility) interval.

First, before doing any analysis, regions of meaningful change can be identified, creating a contextual backdrop to which results can be compared. Based on the information at this US Forest Service webpage (<https://webcam.srs.fs.fed.us/graphs/vis/index.php?wilderness=dollys>), the haze index used in the examples (with units of deciview) can be approximately translated to standard visual range (in miles), and the page explains that a “one unit change in deciview may be noticeable under certain conditions.” A change from 14.0 deciview to 12.5 deciview is translated into a change from about 60 miles visual range to 70 miles (an increase of 10 miles in standard visual range). Let’s suppose that a *change in the mean haze index corresponding to 5 miles in standard visual range over the span of 2 years* is agreed upon by subject matter experts and stakeholders to qualify as meaningful to wilderness character. This translates to a change of about 0.7 deciview over the 2 years for the range of values expected for the park (between about 10 and 15 on the haze index) or a change of 0.35 per year (slope).

Then, one can go through the same exercise about what is *not* meaningful (rather than simply choosing zero). In the future, the WCM language of “stable”, “upward”, and “downward” might be useful to bring in, but for this example we use “meaningful change.” Suppose a 0.5 mile decrease per year in standard visual range is judged to clearly fall short of changing wilderness character. Translating this into a change in the haze index gives a change of about 0.1 deciview per year in the mean haze index. Note that a different change might qualify as meaningful within different ranges (e.g., the difference between 1 mile and 6 miles is probably interpreted much differently than between 65 and 70 miles).

Now, the 0.35 deciview/year and 0.1 deciview/year magnitudes of change can be used to further refine the region of meaningful change to provide practical meaning to statistical results. Anything greater than 0.35 deciview/year (in absolute value) is considered definitely meaningful change and anything under 0.1 deciview/year is considered definitely not meaningful change. There are still limitations to having to set these cutoffs, but this approach still leaves the gray area between -0.35 and -0.1 that represents a continuous transition from meaningful to not meaningful. Allowing room for uncertain gray area and transition is a more realistic way to assess the situation, as opposed to choosing one point that separates meaningful from not-meaningful (or not choosing meaningful at all and relying on a null hypothesis of zero). Drawing the picture as documentation before analysis and for use after analysis is helpful, and the web applet in Appendix A.3 provides an easy way to do this (see Figure 3).



Figure 3 Visual representation of regions specifying meaningful change (cyan) and not meaningful change (green), as well as the unavoidable transition between them.

The picture in Figure 3 provides regions of meaningful change as a backdrop for interpreting the results of the statistical analysis in context (under all the assumptions of the analysis). One way to proceed is using confidence (or compatibility) intervals to convey statistical uncertainty in results, without using them as a hypothesis test or even choosing a set alpha level. Figure 4 provides an example of displaying confidence intervals corresponding to 5 different levels (50%, 75%, 90%, 95%, and 99%) which provides more information than a single interval. Intervals obtained from the Landres et al. (2009) real data (Ex. 1) and modified data (Ex. 2) are provided in reference to the backdrop.

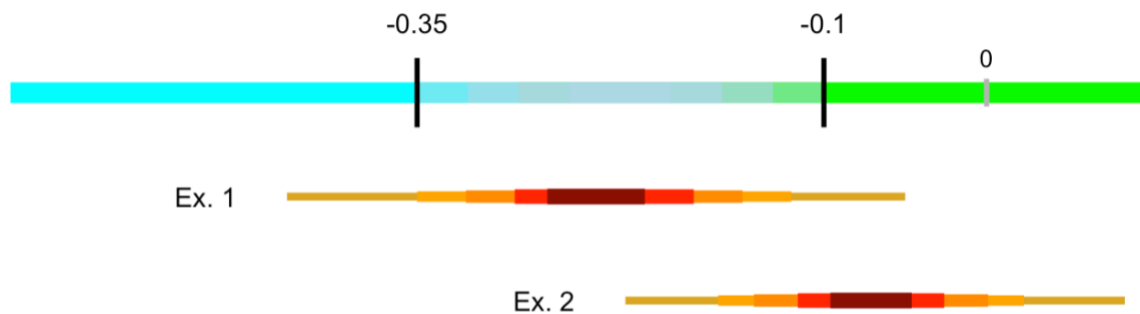


Figure 4 Visual representation of regions specified as practically meaningful (cyan) and not practically meaningful (green), as well as the unavoidable gray area in between, compared to confidence intervals associated with the two examples described in Landres et al. (2009) Appendix B (Ex. 1 is for the real data, and Ex 2 is for the modified data). Instead of choosing one confidence level, the intervals convey the statistical uncertainty through including the 50%, 75%, 90%, 95%, and 99% confidence intervals (under the assumed simple linear regression model).

The task is then to interpret the situation in the context of the problem by comparing the intervals to the regions of meaningful change in the backdrop. Here are some observations (not exhaustive) that might be used to come to a conclusion after also weighing the potential costs of the ultimate decision to classify trend as “upward”, “downward”, or “stable.”

- Ex. 1 results fall mainly in the gray area, not as black-and-white as the “statistical significance” would imply since the confidence intervals do not contain zero.
- The Ex. 1 results are not compatible with a slope of zero, but they are compatible with many other values of the slope that are also not considered practically meaningful, as well as some values that are considered practically meaningful.
- The Ex. 1 intervals cover the gray area and include values in both the practically meaningful and not practically meaningful ranges. Overall, the results for Ex. 1 are largely inconclusive, even though the confidence intervals all exclude zero and according to purely statistical criteria, there is evidence of a “significant trend.” The work of creating the backdrop (the picture in Figure 3) and then interpreting statistical results relative to that backdrop takes more effort and justification than relying on arbitrary summary statistics and thresholds; however, it yields more useful information. The ultimate conclusion should take into account implications of resulting decisions and consequences of erring in either direction (are the consequences more severe for

concluding “declining” when it might be “stable”, or vice-versa?) in the context of the values covered by the interval, rather than a specified type-I error rate.

- Ex. 2 covers a wide range of values that represent changes not thought to be meaningful, but the data and model are also consistent with values that range well into the gray area. Again, the interpretation is not as simple as checking whether 0 is in the 90% confidence interval and saying “no significant trend.” The ultimate decision for Ex. 2 appears more straightforward than for Ex. 1, despite Ex. 1 being associated with “statistical significance” according to the criteria used for the example.

2.5.5 Problems with shifting from percent change and statistically significant change

As discussed in Section 2.5.4.1, for some measures in the *Draft Technical Guide* and *Draft CAAR WCMP*, the threshold for change strategy is to employ a combination of percent change and statistically significant change. The recommendation is to use percent change for the first 5 values of the measure (or after “a sufficient amount of data”), followed by use of statistical significance based on a linear regression model. The combining of the two approaches for a single measure results in inconsistencies that are difficult to justify. This section presents the inconsistencies to promote awareness of the potential limitations and describes how the definition of meaningful change differs among the two options.

2.5.5.1 Multiplicative vs. additive change

A key point to understanding the difference between the two approaches is the distinction between additive change and multiplicative (or percent) change. “Amount of change” can be defined additively or multiplicatively. Multiplicative change describes the change as a multiplicative factor relative to a reference point and does not have units of the measure attached to it (e.g., a doubling, a two-fold increase, a 100% change, etc.). The *Percent Change* threshold in the *Draft Technical Guide* option uses multiplicative change in the form of 3%, 5% or 10% change. Additive change describes the change in terms of the addition or subtraction of a number of units on the scale of the measure (e.g., a one unit increase, a 5 unit decrease, etc.).

A standard linear regression model is defined in terms of *additive* change in the mean of the measure over time (each year is associated with an m unit increase in the mean of the measure). Using a percent-change threshold may implicitly assume that each year the individual value (and maybe mean) increases by some *multiplicative* factor, or just that the percent change for any year should be calculated relative to the baseline measure value (the distinction between these is not clear in the guidance). These are fundamentally different models for a trend; there is a conceptual disconnect between defining the amount of change qualifying as meaningful in terms of multiplicative change for the first 5 observations (potentially 25 years) and then shifting to defining it in terms of additive changes in a mean simply because more than 5 data values are available. This changes not only the threshold for change itself, but a foundational assumption regarding how it is defined.

There may be a situation where this makes sense, but it should certainly not be the default approach and should be backed up by practical knowledge about changes in the system and why the process thought to underly change is expected to change so fundamentally after 5 (or 25) years. For example, for small values of a measure and small percent changes in successive years over time, the trend defined by a multiplicative trend can look approximately linear so that linear regression may be appropriate, but

then the result will be interpreted in terms of an additive change in the mean per year. Again, such a decision should still be justified for a particular measure, expected range of values, etc.

Linear regression applied after natural log transformation of the measure (and/or the time variable) does correspond to multiplicative change on the original scale of the measurement, while being additive on the log scale. Assessing the assumption of linearity using a scatterplot of the measure values can help with this; if it looks linear after transformation, then that is an indication of multiplicative change rather than additive.

It is possible that the allure of the percent change approach as a default, over the additive change, is the ease of specifying a default, automatic value (like 5%) that is applied broadly over different measures and contexts with little expectation for in-depth justification. Specification of additive changes is not as automatic as the change must be attached to the scale of the measure and there is no default value that applies across measures. However, the lack of a default should not exclude the approach from being considered; it is not an inherent disadvantage. While using a default percent change makes the process easier to define and carry out because it removes the need for justification, it presents a limitation in terms of lack of meaningful justification and promotes ignoring the measure-specific context.

There is no single correct choice between additive and multiplicative definition for threshold for change, but the methods then used should be consistent with the choice as justified. It may also depend on what range of values are possible for a given wilderness area, as additive change may be adequate within a smaller range even while multiplicative is needed across larger ranges.

2.5.5.2 Changes in raw values vs. averages

It is important to consider properties of the baseline measure value as compared to future measure values input into the WCMD if the decision about trend really is going to be based on a comparison to the baseline. For example, the baseline measure value might be the value recorded for a particular year (e.g., number of non-native species in 2020). The next value input into the WCMD after 5 years might be another count of species that is directly comparable to the baseline, but it could also be an average count over multiple years (the guidance often refers to 3-year running averages, but never makes it explicit when they will or should be used). Comparing summaries created from different levels of aggregation can impact results and should at least be acknowledged if proposed as part of the process. It clearly affects definition of the threshold for change; the threshold can be defined in terms of raw yearly values (like counts), or averages over multiple years, or the mean change under constraints of a model (such as linear regression). Note that averages are expected to be less variable than raw measure value; the amount of change to qualify as meaningful in an average may be different from that of a raw value. There is no “best” option, but being transparent and explicit about what quantity or summary the threshold applies to is important.

2.6 Statistical significance and default alpha levels

From a statistical perspective, there are multiple concerns related to basing WCM decisions on “statistical significance” and in general, comparing p-values to default alpha levels. The implications of these practices result in lack of transparency related to the actual threshold used, a lack of justification for decisions regarding trend in a measure, and perpetuation of practices and associated language that have been discouraged by statisticians for decades (e.g., see papers such as Greenland et al. (2016),

Amrhein et al. (2016), Higgs (2013), and Gigerenzer (2004)). The following subsections briefly discuss some of these concerns for the purpose of raising awareness and encouraging change in related language and recommendations in future WCM guidance.

2.6.1 Statistical hypothesis tests and decision making

The approach to reject or fail-to-reject a null hypothesis (under the assumptions of a very specific statistical model) can facilitate easy decision-making, but does not necessarily correspond to the practical decision that is desired. In other words, the decision-making process described in a null hypothesis significance test is very limited and restricted in its use; it is not meant to take the place of more complex decision making in broader contexts of science and/or management. It is unfortunate that the ease of applying statistical hypothesis tests to obtain a “yes/no answer” has led to its widespread use in contexts beyond that for which it was designed for. The WCM process can avoid making this mistake by not presenting statistical modeling and hypothesis tests as the preferred way of implementing the WCM process, or at the very least caution against implementing them without adequate justification. A statistical hypothesis test is not capable of blindly answering the question of whether there has been meaningful change in an indicator of wilderness character.

2.6.2 Alpha-levels, Type-I errors, and use of p-values

A p-value may be employed as part of statistical inference without comparing the result to an alpha-level, or desired Type-I error rate. An alpha level refers to the rate of making a Type-I error using a hypothesis test - rejecting the null hypothesis when in fact the null hypothesis is true. In the context of regression and trend, this is translated as rejecting the hypothesis that the slope is equal to zero, when in fact the slope is equal to zero. Alternatively, the alpha level is also connected to the rate of “coverage” – the rate at which a $(1-\alpha)$ confidence interval fails to contain the true value of the slope that generated the data.

In the statements in the previous paragraph, it is important to consider what the “rate” is referring to – it is defined by considering the proportion of Type-I errors (or confidence intervals failing to include the true value) obtained over applying the same statistical method to many hypothetical realizations of data. The hypothetical realizations of data are considered to be generated using the same mechanism that generated the observed data. In practice, we approximate the data generating mechanism with a model and therefore, the direct relevance of “rate” relies on how well the model used captures the actual data generating mechanism. If a design incorporates probability through random sampling or random assignment, then the connection is explicit; different hypothetical random samples or randomizations are the different realizations and the model is well supported by the design. Observational studies with no random sampling or randomization present a more challenging situation and it is often difficult to interpret the meaning of a Type-I error rate within the context of the problem, which calls into question its usefulness. For many measures in the WCM process, it is not clear how relevant the concept of an “error rate” actually is due to lack of randomness in the design and challenges in modeling the data generating process; this calls into question the decision to rely on a particular alpha-level for a crucial part of the WCM process.

The p-value is a summary statistic based on the data, model, and chosen null model and can be interpreted on the continuous scale between 0 and 1 – without reference to an “alpha-level.” Its usefulness depends on the appropriateness of the model and the strength of the connection between

the null model and bigger questions of interest (e.g., has there been meaningful change in wilderness character?) For the context of estimating change using regression for WCM, the p-value adds very little information to the decision process.

2.6.3 A default type-I error rate regardless of measure

The following sentence found in the *Draft Technical Guide* brings up problems related to the suggested approach of comparing p-values to default alpha-levels:

For all measures that use regression in this Guide, an alpha level of 0.1 will be used in determining statistical significance, meaning that there is a 10-percent chance of concluding that there is a significant trend when in reality there is not a trend, or conversely that there is 90-percent confidence or certainty that the trend is real.

There are several common misinterpretations of statistical concepts included in the quote: (1) the statement “there is 90-percent confidence or certainty that the trend is real” is not correct; (2) concluding there “is a significant trend” is more accurately described as rejecting the null hypothesis of a slope of zero under the specified regression model; and (3) default use of a single alpha level for all measures and contexts is not justified. It is also important to remember that classifying a trend as statistically significant does not imply the estimated change is practically meaningful, as discussed previously. These three misinterpretations are discussed in the following paragraphs.

First, statistical “confidence” does not provide information about the chance, or probability, that a trend is “real.” This is a common misinterpretation; it reflects what we wish statistical methods could provide, but not the reality of what they do actual provide. See Greenland et al. (2016) for a succinct overview of this myth, as well as others associated with p-values and confidence intervals. The 90-percent referred to describes the rate at which a 90% confidence interval will include the “true” value if all assumptions are met. This quantity is far from some level of “certainty” that the trend is real. An added complication is the definition of the term “real” in context of using a simple linear regression model. The statement may have been trying to get at something closer to the statistical power, as described further in Section 2.6.3.2.

Regarding the second misinterpretation, it is important to restate “significant trend” in terms of what criteria are used to achieve “significant.” Adopting such a practice avoids overselling the relevance of statistical modeling results to those without enough background in statistics to understand the subtleties contained in the word “significant.” The term also contributes to the ease of confusing a statistically significant trend with a practically significant one.

Finally, recommending use of a default desired Type I error rate (alpha level) across all measures contributes to an expectation that justification of an alpha level need not be measure and context specific. It also assumes that a Type I error rate is meaningful in the WCM context and should be relied on for the crucial decision of whether meaningful change occurred. This is discussed previously in Section 2.6.2 and is also discussed further in the next section.

2.6.3.1 Choice of alpha-level and Landres et al. (2009)

In the context of using the linear regression model, the *Draft Technical Guide* states:

using regression to determine whether a trend is statistically significant requires the user to assign the desired degree of confidence, or certainty, in the results, called the alpha level (there is extensive literature on this topic that is beyond the scope of this Guide.)

There is extensive literature on regression and associated hypothesis tests, but there is also extensive literature explaining misinterpretations of concepts and statistical language, such as the myth that an alpha level controls the certainty in results. As previously discussed, the linear regression model can also be useful without appealing to hypothesis tests, alpha levels, and statistical significance. Results can be obtained without having to categorize them as “statistically significant” or “not statistically significant.” The dichotomy implied is convenient and historically common, but is not typically well justified and is of questionable usefulness in many situations.

The *Draft Technical Guide* also uses the following language to justify the blanket choice of 0.1 for an alpha-level:

This alpha level allows an appropriate balance between the need to catch trends early while maintaining as much statistical rigor as possible in correctly identifying meaningful trends (see Appendix in Landres et al. 2009 for details on the selection of this alpha level and use of regression).

Landres et al (2009) is the Appendix B of the Forest Service *Technical Guide for Monitoring Selected Conditions Related to Wilderness Character*. Given the reliance on Landres et al. 2009 for details and justification for the alpha-level approach, this section provides some additional review of the statistical concepts and language in that document. Landres et al. (2009) use statistical language and concepts in ways that fall into many of the common traps of misusing and/or misinterpreting common statistical summaries and practices. It is my professional opinion that Landres et al. (2009) Appendix B should *not* be referenced in the future as justification for how or why to carry out a statistical analysis for trend in the WCM context.

Statistical rigor, as referred to in the above quote, is a reasonable goal, but should be defined by justification of assumptions and methods relied on for a particular problem; it cannot simply be attained through specification of a desired Type I error rate (the appropriateness of which depends on all the assumptions of the statistical model). Such wording can be misleading, particularly to those stakeholders who do not understand the limitations of statistical methods and who do not have the technical expertise to independently assess the reasonableness of the recommended approach.

A thorough review of Landres et al. (2009) is not provided here, but their recommendations fall five main areas: (1) the trap of relying on a false dichotomy – that results are either significant or not; (2) the trap that statistical significance is analogous with practical significance; (3) over-focusing on Type I error rates, while ignoring Type II error rates; (4) misinterpreting the terms alpha level and confidence level; and (5) ignoring magnitudes of changes while overfocusing on precision. The first three are discussed elsewhere in the review, and issues (4) and (5) are discussed briefly below.

Some of the explanations in the document are simply false from a statistical perspective. For example, the statement that the alpha level means that “when there is a trend in the data, there is a 5-percent chance that this trend is occurring at random, and there is a 95-percent chance that this trend is real.” As previously mentioned, Greenland et al. (2016) may help explain and dispel some of the incorrect

language used. To be fair, such incorrect use of the concept is widespread in science and continues to be perpetuated inadvertently by examples like this set up as references for future documents.

Landres et al. (2009) assume that setting an alpha level is a reasonable way to select “a trigger point or threshold for change.” It goes on to imply that the trigger point is the “alpha level”, which is not the case, and that it “is based solely on managerial and philosophical considerations of how much randomness is acceptable in identifying a significant change.” It is difficult to interpret all parts of this sentence together, but in my opinion the alpha-level appears to be chosen based more on commonly used statistical practices, rather than a connection to management decisions. The discussion suggests an alpha-level is capable of much more than it is from a practical perspective.

I agree with the following statement in Landres et al. (2009): “from a managerial perspective, some changes may in fact be important but they will not be statistically significant.” However, the statement begs the question of why such heavy reliance is then placed on statistical significance (and the assumptions needed to justify its use) in order to judge practical importance. Statistical significance places heavy reliance on precision of the estimated trend and not the amount (or magnitude) of change that qualifies as meaningful. The example provided in Landres et al. (2009) and previously described in detail in Section 2.5.4.4 provides an illustration of this. The focus is on whether the p-value reaches the 0.1 threshold, while magnitude of the estimated slope and values contained in the confidence interval are ignored.

2.6.3.2 Considering Type-I error rate and not Type-II error rate

If it is decided that model assumptions are reasonable and the context lends itself to trying to control a Type I error rate (rate of rejecting the null hypothesis when it should *not* have been rejected), then Type II error rate should also be considered. There is a tradeoff between the two. Type II error rate is more often referred to using its complement, which is power (the rate of rejecting the null hypothesis when it should have been rejected). It is common to refer to these in introductory statistics courses as if they are simple concepts (analogous to false positive and true positive medical test results). However, the concept of “repeat sampling” (analogous to repeating a medical test over and over again) is often difficult to justify and rather far removed from the context of an observational study monitoring a quantity over time. As described above, it is very dependent on the assumption that the statistical model used is a reasonable model for generating alternative (hypothetical) realizations of the observed data, which again gets us back to considering the assumptions of the underlying model.

Assuming the framework is going to be used, the desired rates for Type-I and Type-II errors should be justified based on potential costs or consequences of the errors. This typically requires connecting the results of the hypothesis test to an often informal decision path that may follow from it. That is, it is most useful to consider the practical downstream consequences of each error when setting desired error rates. This is another setting where there is a tendency to just pick default values of 0.05 (or 0.1) for the Type-I error rate and 0.2 for Type-II rate (equivalent to 0.8 for power). It is rare to see meaningful justification for these choices in the context of real decisions that might be made based on the results.

2.7 Making the reference for judging change clear

References to the baseline measure value in *KIW2* and the *Draft Technical Guide* convey that the baseline will serve as the reference value for assessing change in the future (see Section 2.1 for excerpt). However, there are also multiple references in the *Draft Technical Guide* and *Draft GAAR WCMP* suggesting this will not always be the case. One obvious example, as already discussed, is the use of linear regression, but there is also mention of 3-year running averages and wording that does not clearly point back to using the baseline measure value as the reference. Conclusions regarding trend will be sensitive to this distinction and what is actually implemented in practice may depend on how the algorithm is programmed within the WCMD. Park-level users of the WCMD should be very aware if they are judging change from the baseline, from the previous measure value, or from some combination of these.

For example, the following is provided for the Untrammelled Quality indicator *Use of motor vehicles, motorized equipment, or mechanical transport*: “The threshold for meaningful change is any change in the 3-year rolling average measure value. Once there are five measure values, the threshold for meaningful change will switch to regression analysis.” In this case, it does not sound like values are being compared to the baseline measure value, and it is also not clear whether 5 “values” refers to five 3-year averages or five years of annual data. This example illustrates how such recommendations are left open to interpretation and may lead to inconsistencies between what is entered into the WCMD and how trend is then classified by the rules incorporated into the WCMD.

This is integral to the ability to define “trend” and justify a threshold for change relative to it. Explicit definition of “trend” should include detail about reference points, time periods, and natural annual variability. More attention to attempting to define “trend” for specific types of measures in the context of WCM may encourage consistency over time and across wilderness areas. As currently described, conclusions regarding trend may differ substantially depending on what reference and summary are used. This aspect of the process is also crucial to defining a relevant *threshold for change*, choosing a method to estimate change, and consistently implementing the proposed process after a threshold is specified.

Summary The *Draft Technical Guide* states “The status of wilderness character in the baseline assessment is the reference point against which change in wilderness character is assessed and evaluated over time.” This statement is inconsistent with other wording and recommendations and results in a lack of clarity regarding the definition of “trend.”

Recommendations

- Current guidance is unclear about whether the “baseline measure value” will continue to be used as a reference point after more data are collected, as opposed to the previous time period’s value or a summary of values over multiple years (e.g., 3-year average). Improving clarity in recommendations and/or acknowledging the need for flexibility is recommended.
- Users should be aware of what reference value the WCMD will use to make conclusions regarding trend. It may be possible, for example, that the user may expect the comparison to be to the previously recorded value, but it is actually compared to baseline using the automatic algorithm.

2.8 Statistical wording (confidence, certainty, alpha level)

From a statistical perspective, there are many words and phrases that have an everyday meaning as well as a formal statistical meaning. Unfortunately, many formal statistical words are routinely misused and

misunderstood within current scientific culture. In this section, I provide several suggestions or cautions related to specific words with the goal of improving clarity in communication among scientists and more broadly with other stakeholders. The overall concern with the wording is that a non-statistical audience will likely interpret the language as if the methods are capable of more than they are and miss the key point that they rely on many assumptions that may be severely violated in many WCM measure contexts.

- *Determine*: Determine is a common word in scientific communication, and while maybe intended to be loosely interpreted, it implies that the process in question is deterministic – that there is a single and obtainable answer. While parts of the mathematics behind statistical results are deterministic, the process of statistical inference relies on many decisions and assumptions to quantify some sources of uncertainty rather than lead to a correct answer. Often, “determine” should more honestly be replaced by words such as estimate, choose, specify, or justify. In the context of WCM, one can argue that it is impossible to “determine” trend in the measure; instead, it is *estimated* under a set of assumptions and using data available. While it may seem like a purely semantics argument, I have found that people interpret “determine” differently than “estimate.” “Determine” is associated with the existence of a single correct answer, while the “estimate” reflects the methods, assumptions, and uncertainty that go into the process.
- *Trend*: Trend is also a word with both everyday and technical meanings. Even within a statistical context, it does not have a single quantitative meaning. Usually, “trend” is used to capture change in some underlying true state of a system we are trying to measure. There are different ways to quantify and estimate trend and the term should be defined as explicitly as possible within a particular context.
- *Significant*: As described before and elsewhere by others (e.g., Higgs (2013)), the word “significant” carries with it connotations of broad importance that have likely contributed to the overreliance on statistical methods to accomplish decision-making tasks that are beyond that for which they are designed. It is quite possible to describe statistical results without the word significant and it is my opinion that it should be avoided as much as possible. If a hypothesis test is going to be used, replace “statistically significant,” by a reference to the criteria used to declare it; for example, “the p-value of 0.062 is less than 0.10.” This is more transparent and explicit and does not carry the connotations of wide importance to an audience who does not understand the term’s limitations. In general, *KIW2* and the *Draft Technical Guide* do a good job of using “meaningful” instead of “significant” to avoid the confusion between statistically significant and practically meaningful.
- *Trend*: The word trend is another word that can be vague and ambiguous in its meaning. There is a general sense that it means change over time, but in a statistical context it can mean different things to different people depending on discipline and their own history. For some, it is used interchangeably with “slope of linear regression line,” while for others it carries a more general connotation. There is no correct definition, but it is important for clarity and transparency to explicitly define what quantity is being referred to as “trend” in the analysis of data over time.
- In the quote from the *Draft Technical Guide* presented at the beginning of Section 2.6.2, the terms *alpha level*, *confidence*, and *certainty* are used as if they are synonymous – which is not the case. “Alpha-level” is often directly related to a “confidence level,” but not necessarily the word “confidence” which has different interpretations. “Certainty” has varied uses and connotations, and from a statistical perspective neither an alpha level nor confidence level provides a “degree of certainty in the results;” though changing the levels is a choice that affects

the width of confidence intervals (and their coverage of the true value) and the rate of Type-I errors. In a statistical context, it is best to not use the word “confidence” as if it means “certainty.”

2.9 Cautions related to automation

A theme repeated throughout this report is a general caution related to automation of the process across measures, potentially leading to key parts being misused of the process in terms of justification, disconnecting parts of the process from ecological and biological context, and making it harder for stakeholders in general to understand or critique decisions.

Use of a program such as the WCMD has clear benefits in terms of efficiency, but its use to “determine the trends” after receiving limited inputs further contributes to potential problems with automaticity. There may be limited opportunities for user involvement in specific cases where it might be warranted and it may be more difficult to deviate from the national recommendations if the approach does not fit nicely into the rules programmed into the WCMD.

The *Draft Technical Guide* (pg 17) states the following: “Once enough data have been entered to derive trends in all selected measures, the WCMD will determine the trends in the measures, indicators, monitoring questions, qualities, and overall wilderness character for each wilderness by using the assigned thresholds for meaningful change.” For some measures, such as those with categories defined based on meaningful change, the automation of the process is expected to work well. In this case, by the time the WCMD implements the final steps, the decisions have been carefully thought through and justified. However, this is not the case if the “determining” trend is based on statistical analysis and the simple linear regression model, or even on one of the three default “percent change” options.

For those measures with trend defined by the estimated slope from the simple linear regression model, it is assumed that the WCMD would automatically fit the model, obtain the p-value associated with the null model of a slope of zero, and compare to the alpha-level threshold. The problems identified with this approach were previously described, but automation of the method introduces additional problems that should be considered. It makes it difficult to implement the more context-dependent approach described in Section 2.5.

In summary, it is poor statistical practice to rely on fitting a statistical model and carrying out a hypothesis test without carefully checking the appropriateness of the model, but this is encouraged with an automatic process such as described for WCM. Instead, each application of the simple linear regression model should involve assessing assumptions and a more thorough interpretation of results than simply checking a p-value against a threshold. A relatively simple illustration of this being a problem is considering the assumption of linearity. It appears quite possible that the assumption of linearity will be severely violated for many measures, particularly if they are constant over a period of time and then increase or decrease. Also, if one is going to rely on the resulting p-values and confidence intervals, then assumptions should be assessed because their validity depends on justification that key assumptions are not severely violated (e.g., constant variance over time). Therefore, I recommend against letting the WCMD do this automatically if a regression approach is justified for a particular measure.

It is also important to keep the end goal of the WCM process in mind – the use of the nationally agreed upon rules for compiling specific trend classifications into an overall assessment of trend in wilderness character (e.g., classifying each measure as “stable”, “upward”, or “downward”, and then forming conclusions by combining over measures, indicators, and qualities). This end part of the process (carried out automatically within the WCMD) has its own assumptions and additional considerations that are not discussed in detail in this review.

Wilderness character is inherently a qualitative concept and care should be taken in trying to simplify measurement of wilderness character indicators using simple quantitative metrics based on available data. If there is a measure that can more holistically incorporate aspects of wilderness character by defining categories more qualitatively, but ordered according to declining wilderness character, that may be an approach to consider. For example, a category shift could include meeting several criteria meant to more broadly capture wilderness character and could include a combination of qualitative and quantitative criteria.

In general, while there are advantages to automating an approach (e.g., consistency, ease of implementation, less documentation, etc.), automatic adoption of defaults or limited options for any measure also comes with costs in terms of ultimate usefulness of the results. Reducing the prescriptive and automatic nature of the WCM process will require additional effort in the planning phase and may not allow for automatic use of the WCMD. However, it would result in gains relative to transparency of the process and usefulness of the results for stewardship decisions.

2.10 Summary and recommendations

Summary Use of “statistically significant” disconnects the *threshold for change* from a scale and context that can be broadly evaluated and understood. The option discourages meaningful review by stakeholders who lack statistical expertise, instead putting them in a position to blindly trust conclusions. Difficulty in explaining a threshold broadly to stakeholders in a way that can be easily evaluated should be taken as a potential red flag regarding the usefulness of the conclusions.

Recommendations

- If regression is used, the p-value threshold (applied to a particular collection of observations) can be translated into a slope of a line - interpreted as the additive change in the mean of the measure per year over the time period. If regression and/or statistical significance is used, it should be accompanied by explicit translation of the threshold into language understandable by a broad audience, rather than hidden as a p-value cutoff.
- The regression approach does not provide a direct comparison to the baseline measure as described as a goal in the *Draft Technical Guide*; a different approach should be selected if stakeholders are under the impression that conclusions are relative to a baseline and/or this characteristic of using regression should be explicitly described.

Summary A threshold based on “statistically significant” looks on the surface as if it isn’t changing over time (e.g., p-value cutoff remains at 0.10), but the magnitude of change considered meaningful *does* change over time as new observations are added.

Recommendation

- The *threshold for change* specified for a measure should remain consistent over time, unless a change is explicitly justified; this advice is consistent with the recommendation in the *Draft*

Technical Guide. However, this advice runs counter to use of “statistically significant change based on regression.” Resolving this inconsistency will help users and other stakeholders.

3 GAAR-Specific Considerations and Recommendations

As introduced in Section 1.3, Gates of the Arctic National Park and Preserve (GAAR) is a vast, remote, and pristine wilderness. The differences between Alaska wilderness areas and those in the lower 48 states present challenges in terms of implementing national guidance for wilderness character monitoring. This section takes a closer look at some of these challenges in the context of selecting a measure and *threshold for change* for the *Plants* indicator of Natural Quality.

The objective behind monitoring the Natural Quality “is to assess the effects of modern civilization on the integrity of wilderness ecosystems, with a focus on plants, animals, air and water, and ecological processes” (page 40 *Draft Technical Guide*). The monitoring question, indicators, and *strongly encourage measures* are provided in Table 7 (page 40) in the *Draft Technical Guide*, with details about the *Plants* indicator on pages 41-46. The recommendations for measure and threshold for change described may have limited utility for assessing change in wilderness character within a system like GAAR. For example, the vast area makes it difficult to obtain meaningful information from measures based on percentage area covered or infested, and the pristine quality results in baseline measures of zero which creates unique challenges in terms of estimating and defining meaningful change in wilderness character.

The Draft GAAR WCM contains an *Overview of Wilderness Character Monitoring Measures*; this includes a list of the characters, indicators, and draft measures, along with descriptions of the information expected to accompany each measure in a wilderness character assessment report. While the focus of this section is on plants, a summary of all draft measures included in the plan, organized according to stated *threshold for change*, is provided in Section 3.1 for future reference and to provide a broader context for discussions regarding recommended approaches for choosing a *threshold for change*. The current description for *Plants* is then provided in Section 3.2 verbatim from the Draft Gates of the Arctic Wilderness Character Monitoring Plan (*Draft GAAR WCM*).

3.1 All Thresholds for Change options included in Draft GAAR report

To provide GAAR-specific context relative to discussions in Section 2 about approaches for defining a threshold for change, this section provides a list of nine different *Thresholds for Change* found in the current *Draft GAAR WCM*. The criteria proposed as thresholds fall under the three options described in the *Draft Technical Guide* (see Section 2.3.3): *any change*, *percent change*, or *statistical significance based on regression*. However, the Draft GAAR WCM deviates from the national guidance to also include additive change for a couple of measures, and also is more explicit about incorporating 3-year moving (or rolling) averages directly into the criteria. Additionally, it allows for a more qualitative approach based on professional judgement for indicators of National Natural Landmarks.

The following provides a list of unique descriptions of *threshold for change* found across all qualities in the *Draft GAAR WCM*, along with the associated draft measures, the relevant section in the *Draft GAAR WCM*, and the baseline measure value. While reviewing this list, it is useful to consider whether the main comparison should be to the baseline measure value or to previously obtained measure value(s) – a crucial distinction that is vague within the *Draft Technical Guide* and *KIW2*.

“Any change in the index value is considered significant”:

- Natural Quality - *Plants*

- Index of invasive plant species (2-1) (GAAR Baseline = 0)
- Undeveloped Quality – *Presence of Inholdings*
 - Acres of private holding (3-2) (GAAR Baseline = unknown)
- Solitude or Primitive and Unconfined Recreation Quality – *Facilities* that decrease self-reliant recreation
 - Number of authorized constructed recreation features (Section 4-5; GAAR Baseline = 0)
- Solitude or Primitive and Unconfined Recreation Quality – *Management restrictions* on visitor behavior
 - Index of visitor management restrictions (Section 4-7; GAAR Baseline = 4)
- Other Features of Value – Deterioration or loss of integral cultural features
 - Status of integral cultural resources (Section 5-1; GAAR Baseline = TBD)
 - Desired condition for integral cultural features (Section 5-2; GAAR Baseline = TBD)

“The threshold for meaningful change is **any change in the 3-year rolling average** measure value. Once there are five measure values, the threshold for meaningful change will **switch to regression analysis.**”

- Undeveloped Quality – Use of motor vehicles, motorized equipment, or mechanical transport
 - Index of administrative authorizations to use motor vehicles, motorized equipment, or mechanical transport (Section 3-3; GAAR Baseline = unknown)

“The threshold for meaningful change is a **5-percent change in the 3-year rolling average** number of unauthorized actions and persistent installations. Once there are five measure values, the threshold for meaningful change will **switch to regression analysis.** A decrease in the 3-year rolling average beyond the threshold for meaningful change results in an improving trend in this measure.”

- Untrammelled Quality
 - Number of authorized actions and persistent installations designed to manipulate plants, animals, pathogens, soil, water or fire (Section 1-1; GAAR baseline small, possibly zero, even for 3-year rolling average)
 - Number of unauthorized actions and persistent installations that manipulate plants, animals, pathogens, soil, water or fire (Section 1-2; GAAR baseline small, possibly zero, even for 3-year rolling average)

“The threshold for meaningful change is a **5-percent change in the measure value for all selected nonindigenous terrestrial animal species (or backcountry units).** Once there are five measure values, the threshold for meaningful change will **switch to regression analysis.**”

- Natural Quality - *Animals*
 - Index of nonindigenous terrestrial animal species (2-2) (GAAR Baseline = 0)
- Solitude or Primitive and Unconfined Recreation Quality – Remoteness from sights and sounds of human activity *outside* the wilderness
 - Median Noise Free Intervals (NFI) and % Noise free intervals >2 hour long (Section 4-3; GAAR Baseline = see baseline for values from individual units and different sites (acoustic inventory or ranger patrol) range from 0.3 to 9.0.)

“A **10-percent change in the measure value or number of encounters or visitors.** Once there are five measure values, the **threshold for meaningful change will switch to regression analysis, and statistical significance** will determine the trend in the measure”

- Solitude or Primitive and Unconfined Recreation Quality – Remoteness from sights and sounds of human activity *inside* the wilderness

- Number of visitors transported by CUAs (Section 4-1; GAAR Baseline = 962)
- Visitor encounters per backcountry unit (Section 4-2; GAAR Baseline = <1 per day: 0.94 in 2017, 0.75 in 2018, 0.35 in 2019)

“A 3-percent or greater change in the measure value will result in a change in trend for this measure. Once there are five measure values, the threshold for meaningful change will **switch to regression analysis, and statistical significance will determine the trend** in the measure. A decrease in the development measure value corresponds with an improving trend.”

- Undeveloped Quality – Presence of non-recreational structures, installments, and developments
 - Index of authorized non-recreational structures, installations, and other administrative physical developments (Section 3-1; GAAR Baseline = unknown)
- Solitude or Primitive and Unconfined Recreation Quality – Facilities that decrease self-reliant recreation
 - Miles of developed trails (Section 4-6; GAAR Baseline = 0)

Specified **additive change** in the value of the measure (e.g., “Any change of 0.5 kg/ha/yr or more in either direction from the baseline data value is considered significant.”)

- Natural Quality – Air and Water
 - Nitrogen Deposition (Section 2-3; GAAR Baseline = 0.2 kg/ha/yr)
 - Sulfur Deposition (Section 2-4; GAAR Baseline = 0.2 kg/ha/yr)

P-value threshold of 0.05 based on linear regression analysis: “Linear regression analysis, $\alpha = 0.05$, for all data available is used to determine if there is a significant trend in the measure.”

- Natural Quality - *Ecological Processes*
 - Extent of Landscape Connectivity (Section 2-5; GAAR Baseline = 99.92%)

“The threshold for meaningful change should be **determined by the park**, depending on the specificity of condition data available when the NNL was designated. If condition status at the time of designation was more qualitative in nature, **professional judgement** should be used to determine significant change. **Regardless of approach, thresholds and determinations should be clearly documented so as to ensure future replicability.** Improvements to the condition of the NNL, that exceed the park-determined threshold, result in an improving trend in this measure.”

- Other Features of Value – *Deterioration or loss of integral cultural features*
 - Significant threats to National Natural Landmarks (Section 5-3; GAAR Baseline = 1)

3.2 Natural Quality – Plants from Draft GAAR WCMP

The proposed measure for the Plants indicator of the Natural Quality character is the total count of non-native plant species, with the “any change” threshold for change. The measure is described in Section 2-1 of the *Draft GAAR WCMP* (currently pages 36-39); relevant excerpts are included verbatim below for ease of reference. Much of the language in terms of defining meaningful change and choice of thresholds comes directly from pages 19-22 of the *Draft Technical Guide*.

2-1 Index of Invasive Plant Species

Wilderness character monitoring baseline	Trend since wilderness designation	
Baseline data value: 0 species		Condition: Good
Years of data: 2014-2017 (four calendar years)		Trend: Stable
		Confidence: Low

Background & Context

In the Gates of the Arctic Wilderness, the primary areas of concern for the introduction of non-native species are travel corridors, such as the John River, aircraft landing sites, and other locations that receive frequent use. In addition, due to the large size of the Gates of the Arctic Wilderness, there are concerns about the staff's capacity to detect invasive species before they become widespread. At this time, no non-native or exotic plant species are known to be present in the Gates of the Arctic Wilderness.

Measure Description & Collection Protocol

This measure monitors the number of non-native or exotic plant species documented within the Gates of the Arctic Wilderness. All species considered non-native or exotic by the Alaska Exotic Plant Management Team are counted within this measure. This measure monitors the *total* number of known non-native species present in the Gates of the Arctic Wilderness, not only new discoveries. For instance, if pineapple weed is discovered in 2020, it would contribute a value of "1" to the total value entered into the WCMD every entry thereafter unless pineapple weed is eradicated. As of 2020, there are no known non-native or exotic plant species present in the Gates of the Arctic Wilderness. However, there is potential for non-native plant species to be introduced or found. A list of invasive plant species common to Alaska is available in Table XX, along with the invasiveness ranking of each species on a scale of 1-100, where 100 is the most invasive and 1 is the least, as evaluated by the Alaska Exotic Plant Management Team. The total number of known non-native plant species present in the Gates of the Arctic Wilderness is entered into the WCMD every five years. Over time, an increase in the number of non-native plant species would indicate a downward trend for this measure.

Data Source

Alaska Exotic Plant Management Team data *and* through Alaska Exotic Plants Information Clearinghouse (AKEPIC), available online at: <http://aknhp.uaa.alaska.edu/maps-js/integrated-map/akepic.php> (accessed August 2014). Both sources should be checked.

Data Adequacy

Low – data quantity and quality are low and medium, respectively, for this measure because data are drawn from the research and datasets of the Alaska Exotic Plant Management Team, which only visit the Gates of the Arctic Wilderness sporadically. However, data that is collected is of good quality.

Frequency

Every 5 calendar years data for this measure will be entered in to the WCMD.

Threshold for Change

Any change in the index value is considered significant. A decrease in this measure's index value, whether due to fewer documented invasive plant species or to species with lower impact ranks, corresponds with an improving trend in this measure.

3.3 Challenges for GAAR

National guidance for the *Plants* indicator, and likely others, does not have seamless application to parks like GAAR that differ substantially from parks in the lower 48 states. Three characteristics of GAAR leading to challenges are (1) the vast area covered, (2) the remoteness and lack of access, and (3) the relative pristine nature. Ironically, these aspects are fundamentally linked to wilderness character, but also make it difficult to assess changes in wilderness character over time according to national guidance using approaches that may be successful in smaller, less remote, less pristine areas.

3.3.1 Data and sampling

The vast area of GAAR introduces challenges in the sampling and monitoring of plants, given that only a very small area can be physically surveyed or sampled. Potential measures are limited, relative to smaller more accessible parks, given the infeasibility of widespread or frequent sampling. Data available from sources such as remote sensing or satellite imagery may be more important for GAAR and other Alaska parks because of the challenges with physical sampling.

Beyond sampling challenges, the vast area also results in summaries based on a “percent area covered” of non-native plants to be of limited usefulness and difficult to calculate with reasonable accuracy. This is important because this type of summary is often relevant for smaller parks with greater relative areas of infestation. Changes in area covered are often summarized in terms of a multiplicative (or percent increase) and for a park like GAAR, the area covered by invasive species may be very small relative to the entire area of the park (e.g., less than 1 acre out of 8.4 million acres). In this context, a change from 0.5 acres to 1 acre (a doubling) can lead to the appearance of a concerning increase in coverage of invasive plants if not considered relative to the total area of a park and the broader context of changes to wilderness character.

As described in the KIW2 and the *Draft Technical Guide*, the WCM process is designed to use data already being collected for other monitoring efforts or objectives; it is not designed to have the resources to collect data for its own objectives. This places additional constraints on possible measures and requires an understanding of historic and current data collection efforts in the park (e.g., Inventory and Monitory Program, Arctic Network, and the Alaska Exotic Plant Management Team (EPMT)).

WCM monitoring objectives may be related to those for other monitoring programs (e.g., an interest in tracking non-native species), but also have potentially important differences. Fitting existing data into meeting the goal of WCM monitoring is in many ways a different exercise than starting from the monitoring goal or objective and designing a data collection and sampling protocol to meet that objective. There are many advantages to using existing data if relevant, but this constraint to the process should be kept in mind throughout the planning process and any limitations arising from it should be explicitly documented. The extent to which the measure based on available data captures aspects of wilderness character and associated changes should be carefully assessed.

Beyond choice of measure and its data source, analysis and decision making set up for another objective (such as monitoring change in vital signs for IMP) may not be directly relevant for monitoring meaningful changes in wilderness character. Therefore, even if data collected for another program are used, the process of deciding upon an analysis and decision-making criteria (e.g., defining a threshold for change) should be undertaken from the perspective of WCM and not simply adopted from the other program using the data. The *threshold for change* should be considered at the same time a potential measure is considered; the use of both together defines the current WCM process and interrogating both simultaneously can lead to important insights regarding the measure that are helpful to discover early in the WCM planning process.

3.3.2 Pristine baseline of no non-native plant species

Unlike most areas in the lower 48 states, GAAR is starting from a baseline of no known non-native plant species. This has multiple implications in terms of choice of a quantitative measure and defining meaningful change in that measure. Measures used in other areas, such as indices of invasive plants, may not be immediately useful or relevant in GAAR. This challenge is discussed in greater detail later in this section relative to potential measures, and the importance of considering measure and threshold for change simultaneously is reiterated. GAAR is also encouraged to consider multi-stage approaches that allow for a change in approach after detection of non-native species, or possibly the use of categories defined more qualitatively based on multiple sources of information and professional expertise.

3.4 Measures proposed for Plants in WCM documentation

There are two measures referred to in either the *Draft GAAR WCMP* or the *Draft Technical Guide*: (1) an *Index of Invasive plants* (the national level *strongly encouraged measure* in the *Draft Technical Guide*) and (2) a count of non-native species (the measure currently recommended in the *Draft GAAR WCMP*). Both of these options has advantages and disadvantages relative to characteristics of GAAR, some of which are identified in this section as a starting point to motivate further discussion and consideration of alternatives. Incorporation of threshold for change into discussions regarding choice of measures is also important.

3.4.1 Index of Invasive Plant Species

The *strongly encouraged measure* in the *Draft Technical Guide* is an index combining multiple types or sources of information into a single quantitative measure, called the *Index of Invasive Plant Species*. It takes categorical sub-measures (described in detail later) and assigns numeric “scores” that are summed to get a score for a species and then summed over species to get an overall score for a wilderness area. While the proposed measure may be well-justified for some wilderness areas, there are challenges to implementing the measure directly for GAAR at this point in time.

As is the case with most indices of this type, the scale of the index is not inherently meaningful without connecting it to a context; for example, it is not clear what a one unit change in the index represents. Therefore, such quantitative indices require caution in terms of interpreting changes in the score using common approaches for summarizing change in quantitative measures that do have inherent or broadly understood meaning (e.g., a change in area, number of species, etc.). The usefulness of such indices in WCM process depends heavily on the ability to interpret changes in the score relative to wilderness character, which relates directly to how to specify a *threshold for change* based on the measure.

3.4.1.1 Calculating and interpreting the Index

The details of the proposed *Index of Invasive Species* (the *Index*) are summarized here, with a full description provided on pages 42-45 of the *Draft Technical Guide*.

Calculating the *Index*:

- The *Index* combines information about (1) presence/absence (a list of detected invasive species); (2) potential impact using the I-rank (US Invasive Species Impact Ranks conveyed through integer scores: 0, 1, 2, or 3); (3) observed impact (using a summary of overall trend in total area of the

wilderness infested conveyed through integer scores: 0, 1, 2, or 3); and (4) whether the invasive species co-occurs with a federally listed animal or plant species (0 or 1).

- The minimum score for a species with trend in infestation classified as “declining” or “unknown” is 1 if an I-rank is available, and 0 if an I-rank is not available. The maximum score for a species is 7, corresponding to a species with an I-rank classified as “high” (score of 3), trend in infestation classified as “increasing rapidly” (score of 3), and found in an area with federally listed plant or animal species (score of 1).
- The overall index score for the wilderness area is obtained by summing the scores from all species, ranging in possible values from 0 to the number of species on the list multiplied by 7.

Interpreting change in the *Index*:

- The threshold for change is described as “any change in the index value is considered significant.”
- A change in the index may reflect a change in the area infested *and/or* a change in the number of invasive species. The “any change” threshold does not distinguish between these in terms of a conclusion about trend in wilderness character. This is an example of a simple and automatic threshold that may be disconnected from the baseline measure value and larger context surrounding the area monitored.
- Conclusions regarding trend do not take into account the magnitude of the baseline (or previous) value. For example, consider three different changes: 0 to 1, 0 to 7, and 15 to 16. As is, these three scenarios would lead to the same conclusion of “declining trend” in wilderness character for the indicator, despite potentially representing very different practical scenarios. This illustrates the importance of considering the measure and threshold together. In this case, the *Index* includes valuable information about potential and observed impacts, but when used with the “any change” threshold the depth of information is not reflected in the ultimate conclusion.
- The *Draft Technical Guide* is unclear about whether the reference for assessing change is meant to always be the baseline value, or if it is expected to shift over time. For example, if over the first 5 years the index changes from 0 to 10 (“declining trend”) and then decreases from 10 to 4 in the subsequent 5-year period, does that deserve a “declining trend” (relative to baseline) or an “improving trend” (relative to the previous time period)? Is there a recommendation and process for including and conveying both?

3.4.1.2 Advantages for GAAR

Advantages of the *Index of Invasive Plant Species* for GAAR include:

- It provides consistency in measures with national guidance and other wilderness areas.
- The measure would serve to document the shift from no known invasive plant species to the detection of species in the area (assuming information is available for an I-rank and for assigning a trend in infestation area).
- When/if an invasive species is detected, the index value provides information about the potential for impact in the park (assuming an invasiveness rank is available for the species).
- The index may capture important information about human spread of invasive species (e.g., airstrips, ATVS, backpackers, etc.), assuming adequate data are, or can be, available.

3.4.1.3 Disadvantages for GAAR

Disadvantages of the *Index* for GAAR include:

- The information included in the *Index* beyond presence/absence is not immediately useful given GAAR's baseline of no known invasive species.
- The "any change" threshold may document detection of an invasive species, but will not be sensitive to what species is found even though a score based on the species' invasiveness rank is included in the *Index*.
- The *Index* uses a summary of trend in amount of wilderness area infested, which may be difficult to obtain given the vast areas of AK parks, relatively small areas that would be initially occupied, and constraints on the amount of area that can be sampled. This part of the measure would likely require additional definitions and guidance to make it useful for AK.
- If a detected species has an unknown I-rank and unknown trend in infestation, the *Index* value will be zero and will fail to record it as a change by the stated criterion. The extent of this problem depends on availability of useful I-ranks and extent of infestation data, both of which may present greater challenges in Alaska than the lower 48 states.
- I-ranks may not be available for species present or may not be applicable to Alaska ecosystems if developed for the lower 48 (e.g., a species with a "high" potential for negative impact in a lower 48 park may not have as severe of potential in Alaska, or vice versa). The database will no longer be updated and it is not clear how well the current I-ranks cover species expected to be a problem in Alaska or whether they reflect their potential for infestation in the Alaska ecosystem. Therefore, developing I-ranks may be needed for the measure to be useful and to avoid automatically assigning 0's to species with unknown I-ranks.
- If values are always compared to the baseline measure value of zero, it will never be possible for GAAR to have an "improving trend" conclusion. This is not necessarily a disadvantage, but should be considered relative to the WCM process and management decisions.
- The *Index* only includes species classified as invasive (potential for economic or environmental harm) rather than all non-native species. Alternatively, this can be an advantage if the park does not consider all non-native species to be a potential threat to wilderness character.
- The *Index* does not account for weedy native species whose distribution over space may change in response to climate change, potentially affecting wilderness character. More generally, it does not account for changes in the distribution of vegetation cover classes over the landscape (e.g., increase cover by shrubs or spruce trees moving into areas that were previously tundra).

3.4.2 Count of total non-native species

At the time of this report, the measure for the *Plants* indicator in the *Draft GAAR WCMP* is the count of total non-native species and the stated *threshold for change* is *any change*. These two choices together define meaningful change in wilderness character as the detection of any non-native species. Values of the measure would be recorded every 5 years, starting from a baseline of zero – no known non-native species.

The specified measure and its threshold for change are relatively simple to justify, and assuming non-native species would first show up in places already being sampled or visited by humans who would recognize the species, it may be realistic in terms of collecting the needed information.

The advantages and disadvantages of this measure overlap with those of the Index of Invasive Species, though there are some important differences as well.

3.4.2.1 Advantages for GAAR

Advantages of the count of non-native species for GAAR include:

- Ability to capture a shift from no known species to presence of non-native species.
- Includes all non-native species, in addition to invasive species (the importance of this property depends on how the additional species are connected to wilderness character).
- An I-rank is not needed for each species.
- A summary of trend in the area of the wilderness infested is not needed.
- Will be greater than zero if any non-native species are detected, thus avoiding the potential problem of the *Index* remaining zero if a detected species has an unknown I-rank.
- Expands the list of included species to all non-natives, rather than just those classified as invasive. Therefore, this measure is more sensitive in terms of capturing a change in species found. The importance of that sensitivity depends on how the additional species are connected to wilderness character.

3.4.2.2 Disadvantages for GAAR

Disadvantages of the count of non-native species for GAAR include:

- Does not include the species-specific information potentially captured in the *Index*, such as potential for and observed extent of infestation. Instead, all non-native species are treated equally. However, when both are used with *any change*, there may be little difference between this simpler measure and the *Index* in terms of conclusions about trend in wilderness character.
- As with the *Index*, if the reference is always the baseline measure value, then “improving trend” will never be an option in GAAR (this may not be a disadvantage).
- As with the *Index*, it does not capture infestations of weedy native species or changes in spatial distribution of other species (e.g., an increase in shrubs or spread of spruce into tundra).

3.5 Alternative approaches

A few alternative approaches to consider are described in this section. A theme to these approaches is creativity and flexibility in definitions and considering the measure and threshold simultaneously in order to successfully capture meaningful change in wilderness character. Alternatives also encourage combining multiple sources of information, including professional judgement, into one or two measures, rather than relying only on quantitative data immediately available.

3.5.1 Categories defined according to meaningful change

The “Threshold for Change” description on page 19 of the *Draft Technical Guide* refers to an *Index of Invasive Plant Species* measure that is different from the one presented as a strongly encouraged measure and described in Section 3.4.1. This different index is presented for illustration within the description of the *any change* threshold option and it is worth describing here as it provides an example of defining a measure using categories that are explicitly defined such that a shift in category represents meaningful change in wilderness character.

The wording in the *Draft Technical Guide* is as follows: “For example, when professional judgment is used for the *Index of invasive plant species* measure, the any-change threshold is used because any

change between the defined categories (non, low-, moderate-, or high-estimated percentage of wilderness occupied by nonindigenous plants) would be, by definition, a meaningful change.”

This provides a straightforward example of simultaneously defining the measure and threshold for change relative to the WCM objectives, and of using a more qualitative approach based on professional judgement to accomplish this. In this situation, the measure described takes the emphasis off the quantitative details included in the measures described previously. It uses professional judgment to simultaneously define the measure based on the *any change* threshold applied to categories. This does not mean any change in percent wilderness occupied is considered meaningful, only that a change in *category* is meaningful – because the categories are created so that any shift between categories over time is considering meaningful change.

It can be tempting to rely on common quantitative techniques to estimate change and judge the magnitude (such as “regression and statistical significance”). However, strategies such as creation of categories tied to meaningful change, from a more qualitative perspective, may better accomplish the goals of monitoring the complex concept of wilderness character, particularly when ideal quantitative measures are difficult, if not impossible, to obtain. Such categorical measures can be based on multiple sources of information – including data from other monitoring programs and professional judgment of those with knowledge and expertise of GAAR.

Current scientific culture tends to rely heavily on quantitative and statistical methods for monitoring and detecting change, even of very complex and qualitative properties such as wilderness character. Statistical methods are often assumed to provide “rigor” and “objectivity” to the approach. However, this perception often comes at the cost of divorcing the methods from the broader context and decision framework, and reducing transparency due to general lack of understanding of the methods and their assumptions. The benefits of using statistical methods in a context such as WCM depend on the strength of the connection between the underlying goal and the assumptions underlying the method. Defaulting to using statistical methods can lead to unnecessarily restricting options for monitoring that may be more effective in accomplishing the goal. Incorporating information in the form of professional judgement and knowledge from expertise that cannot be easily captured in a quantitative measure is often labeled as “subjective” and less desirable than quantitative statistical approaches. The pros and cons should be considered in the context of each indicator, and a middle ground approach such as the definition of categories based on both quantitative and qualitative information may also be an option.

3.5.2 A multi-stage or adaptive approach

A possible approach to addressing the challenge of monitoring during an underlying change in states (e.g., no non-native species to monitoring extent of infestation of species) is to adopt a multi-phase or adaptive approach, where the first stage is focused on presence/absence (or detection) and the second phase is focused on monitoring infestation. This approach may be valuable for other potential measures, such as changes in spatial distribution of species over time as well. For example, a measure could focus on whether spruce is detected in an area that has been historically tundra, and then shift to a strategy for monitoring the change in extent over time. A middle phase could also be added that focused on the potential for severity of the problem – lying between detection of a potential problem and monitoring the actual severity (e.g., use of invasiveness ranks captures this type of idea).

Overall, variations on such a multi-phase approach could be developed using professional expertise and understanding of the GAAR ecosystem and potential changes in the future related to wilderness

character. New measures and thresholds may be defined after pre-specified criteria are met, with the new measure and threshold potentially dependent on the state of the system at the time (e.g., what species were found or suspected to be found in the near future?)

3.5.3 Beyond invasive species

The national guidance emphasizes infestation of invasive species, an emphasis that makes sense for the lower 48 states. For Alaska, other potential changes in plant communities may be of higher priority than the potential for infestation of invasive species. For example, in a discussion with Peter Neitlich, he highlighted the potential importance of emphasizing changes in the distribution of native species over the landscape in Alaska (e.g., increase in shrubs and spruce growing in areas previously tundra), as is currently happening and expected to continue due to climate change. The default choice of focusing on invasive plant species would miss such important aspects of changes in the Alaska system due to climate change.

3.6 Potentially useful questions for developing the GAAR WCM

The following questions may provide a starting place for more in depth discussions about choice of measure and threshold:

1. What is considered the biggest threat to wilderness character in terms of the *Plants* indicator for GAAR?
2. What are the potential negative aspects of missing or ignoring information about invasive species, or vice-versa, about relative changes in spatial distribution and extent of native species?
3. Is it worth using two different measures (one for invasive species and one for distribution of native species)? Note: If two measures are created, the default is to give them equal weight in the WCM process, and therefore this should be considered in the process of measure and threshold development.
4. Is it worth creating a measure that captures and weights information from both non-native and native species?
5. Are there other types of information or aspects related to the *Plants* indicator that should also be considered for GAAR wilderness character?
6. What data are or will be available to provide information and support each measure? Less formal data based on professional expertise and/or observation may also be considered.
7. What process will be used to develop and justify a *threshold for change* simultaneously with each measure? It is important to move the discussion of what qualifies as meaningful change earlier in the process and to treat measure and threshold together – as they work together to translate information into a conclusion about trend in the indicator of wilderness character. Questions such as following can be helpful:
 - What are the realistic values this potential measure could take on within GAAR in the next 5 years? 10 years? 25 years? 50 years?
 - Is the amount of change that could realistically happen within the time ranges considered meaningful in terms of wilderness character?
 - Is any change of the same magnitude (e.g., a 1-unit change) judged to represent the same amount of change to wilderness character? (e.g., 0 to 1 species as compared to 3 to 4 species)
 - How would you convince another stakeholder why a particular change qualifies as meaningful?

8. How might resulting conclusions about trend (based on the measure and threshold) potentially inform management and policy decisions? How sensitive should the threshold be to change? That is, should it serve as a red flag for change that may be coming or should it sound an alarm after meaningful change has already occurred? Note: this idea is somewhat analogous to the logic appealed to in the use of Type I and Type II errors in the context of statistical tests, but considered from a broader and more holistic view beyond the narrow perspective of a statistical model and an associated null hypothesis significance test.

3.7 Additional recommendations and considerations

Additional recommendations and considerations not included in the previous sections are included here, first those specific to Plants in GAAR and second more general recommendations that apply across measures. There is some overlap with the general recommendations provided in Section 2, but they are not all reiterated here.

3.7.1 *Plants-specific recommendations*

- The possible measure of % area covered by shrubs (advancement of woody vegetation) was suggested by Peter Neitlich, though a measure based on this idea need not be a quantitative summary with a simple quantitative threshold (like change in cover by 1%). It could be as simple as shrubs becoming the main cover in a particular area, or spruce being found in a particular area of tundra. With continued climate change, trend in measure based on the spatial distribution of species is expected to be “stable” or “declining” but not improve.
- Based on the previous statements, the new measure should take into account not only presence/absence of non-native plants, but also ultimately capture characteristics of the species present and ultimately the severity of an infestation (or shift in species) when relevant. To capture potential scenarios for how the system may change over time, it will be important to allow the measure and threshold to evolve with a changing system, when justified - finding a balance between rigidly adopting common or recommended measures and creating a system that changes too often without justification.
- If formal statistical analysis is to be used in the process of assessing change, then the quantitative properties of the measure should be specifically assessed using the Quantitative Property Questions in Appendix A before selecting an approach, as well as consulting a statistician.

3.7.2 General recommendations

- While it may be attractive to boil each decision down to the comparison of a quantitative measure against a pre-specified threshold, this approach may not lead to the most effective monitoring strategy, particularly when relying on automatic and default methods. Use of statistical methods and common, though arbitrary, statistical thresholds should not be interpreted as having greater rigor than other approaches. For some indicators and measures, use of statistical methods to estimate change may be justified, assuming the model is connected to the biological or ecological context of the measure and the connection to any threshold for change.
- As described in Section 2, statistical methods cannot, by themselves, be expected to identify a magnitude that qualifies as meaningful change. Defaulting to the use of statistical methods and

thresholds often takes away some of the burden of justifying decisions and sends a message of “objectivity” – but the perception of objectivity often comes at the expense of divorcing the decision framework from the meaningful context of the problem. Statistical inference has serious limitations in the WCM context and greater sophistication in terms of statistical modeling does not imply increased scientific rigor or reasonableness in decision making. Greater sophistication in terms of statistical modeling does not imply increased scientific rigor or increased reasonableness in decision making.

- Wilderness character is inherently a qualitative concept and capturing changes in some indicators of it with one or two quantitative metrics based on available data may be challenging. It may be beneficial to allow for incorporation of sources of information beyond traditional “data,” such as professional expertise and knowledge of the ecosystem. As discussed earlier, a set of categories ordered according to declining wilderness character may be an approach to consider; a category shift could require meeting several criteria meant to more broadly capture wilderness character and could include a combination of qualitative and quantitative criteria.
- Potential measures lie on a continuum from those that are quantitative measurements of a single characteristic (count of species, proportion of total area covered, etc.) to qualitative categories meant to consolidate information from many characteristics. In the middle of these two options are indices created by combining quantitative and qualitative information into a numeric index – like the strongly encouraged measure Index of Invasive Plants. It may be useful to keep the continuum in mind to avoid unnecessarily restricting potential measures to one part of the continuum.
- Alaska parks may be best served by taking advantage of flexibility in the process to define a measure and threshold for change that makes sense for the landscape and other context. This will require more work in the process of creating a WCMP, but less work and justification later in the process to support conclusions and decisions.
- An effective process for developing or deciding on a measure and threshold may benefit from stepping back in the process and more holistically envisioning what a change in wilderness character might look like for a particular indicator, like Plants – initially ignoring available data to avoid unintentionally constraining the options and the process. This is expected to look different for each park and can lead to taking advantage of the flexibility built into the process.
- Related to the above points, considering big questions related to meaningful change may help focus discussion about measures, as opposed to leaving the threshold for change as a last part of the development as is currently implied in the Draft Technical Guide.
 - Questions such as “What might be some scenarios (described qualitatively) that would represent the smallest change in the plant community that would be considered meaningful change in the Plants Natural Quality of wilderness character?” may be useful in the exercise. Envisioning such hypothetical scenarios can be done before a list of potential measures and available data are highlighted; that is, consider what the system might look like after what would be considered a change to wilderness character in terms of *Plants*. Scenarios can be described in as much detail as possible, and then from the scenario descriptions key aspects may be identified and incorporated into potential measures after assessing what information is reasonable to collect via traditional data sources as well as more qualitative expertise and understanding of the system.
 - It may also be helpful to assess the agreement on such scenarios by people with knowledge of the GAAR ecosystem. Is there general agreement? Where do people differ on their opinions about what changes in Plants may qualify as meaningful change in wilderness character?

- Leveraging the flexibility in the system to develop unique measures for a park will require more written justification, but that expectation may also help strengthen the development process and ultimately the conclusions of the program.

References

- Amrhein, Valentin & Greenland, Sander & McShane, Blake. (2019). Scientists rise up against statistical significance. *Nature*. 567, pp305-307. 10.1038/d41586-019-00857-9.
- Cundill, G., Cumming, G.S., Biggs D., & Fabricius, C. (2010) Soft systems thinking and social learning for adaptive management. *Conservation Biology* 26(1), pp 13-20.
- Gigerenzer, G. (2004) Mindless statistics. *The Journal of Socio-Economics*. 33(5), pp. 587-606. 1053-5357, <https://doi.org/10.1016/j.socec.2004.09.033>.
- Greenland, S., Senn, S.J., Rothman, K.J. *et al.* (2016). Statistical tests, *P*-values, confidence intervals, and power: a guide to misinterpretations. *Eur J Epidemiol* **31**, 337–350 (2016). <https://doi.org/10.1007/s10654-016-0149-3> [Note: Also found as supplementary material to the *The American Statistician's* on-line supplement issue related to the American Statistical Association's Statement of *P*-values <https://amstat.tandfonline.com/doi/full/10.1080/00031305.2016.1154108>]
- Higgs, M.D. (2013). Do we really need the s-word? *American Scientist*, 101(1), pp. 6-9. DOI: 10.1511/2013.100.6
- Landres, P., Barns, C., Boutcher, S., Devine, T., Dratch, P., Lindholm, A., Merigliano, L., Roeper, N., Simpson, E. (2015). *Keeping it Wild 2: An Updated Interagency Strategy to Monitor Trends in Wilderness Character Across the National Wilderness Preservation System (KIW2)*. Gen. Tech. Rep. RMRS-GTR-340. Fort Collins, CO: U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station. 114 p. [KIW2 (2015)]
- Landres, P., Boutcher, S., Blett, T., et al. (2009). Appendix B: Statistical Analysis of Trends in the Measures. Appendix to the *Technical Guide for Monitoring Selected Conditions Related to Wilderness Character*. Gen. Tech. Rep. WO-80. Washington, DC: US Department of Agriculture, Forest Service, US Government Printing Office. 293 p.
- National Park Service (2020). *Draft NPS Wilderness Character Monitoring Technical Guide*. [Draft NPS WCM Technical Guide_11-12-20.pdf]
- Pace, K. (2020). *Draft Gates of the Arctic Wilderness Character Monitoring Plan*. Gates of the Arctic National Park and Preserve, Natural Resource Report NPS/XXXX/NRR-2020/XXX. [2020 GAAR WCM_DraftReport_KKP_2.6.20]
- R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>
- Wickham, H., François, R., Henry, L. and Müller, K. (2020). dplyr: A Grammar of Data Manipulation. R package version 1.0.2. <https://CRAN.R-project.org/package=dplyr>
- Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. <https://ggplot2.tidyverse.org>

Wickham, H. and Bryan, J. (2019). readxl: Read Excel Files. R package version 1.3.1. <https://CRAN.R-project.org/package=readxl>

Xie, Yihui (2019). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.25.

Xie, Yihui (2015) *Dynamic Documents with R and knitr. 2nd edition*. Chapman and Hall/CRC. ISBN 978-1498716963

Xie, Yihui (2014) knitr: A Comprehensive Tool for Reproducible Research in R. In Stodden, Leisch and Peng, editors, *Implementing Reproducible Computational Research*. Chapman and Hall/CRC. ISBN 978-1466561595

APPENDIX A – Supporting Resources and Documentation

A.1 Assessing quantitative properties of measures for a WCM

How a quantitative measure is used in the WCM process should depend on properties of the underlying data and the measure itself. Explicitly considering such properties is important to justifying adoption of a measure, choice of a *threshold for change*, and any statistical modeling that may be proposed as part of the WCM process for that measure. This section provides some questions to help guide the process of identifying and describing important properties. While not exhaustive, the list is intended to support a team with quantitative scientists in the early steps of creating a WCM plan. It encourages time to step back and ask critical questions in the “select measures and thresholds” step of the process before adopting default or recommended approaches. For example, the recommendations in the *Draft Technical Guide* to use regression after five values are obtained or for measures that are “complex” may inadvertently lead users to skip this crucial step in the statistical modeling process.

This set of questions represents a starting point that could be refined and added to if adopted as part of the planning process. A less technical version of the questions could also be developed in future work; it would be targeted to people with less quantitative background. However, reliance on quantitative measures and statistical analysis ultimately should require involvement of quantitative scientists in the process.

Additive or multiplicative change?

1. What is the *smallest* non-zero change that could realistically happen in 1 year? ____ 5 years? ____
 - When you answered the question, did you think in terms of multiplying (e.g., a percent change, an x-fold increase, etc.) or did you think in terms of adding or subtracting (an x unit change)?
2. What is the *largest* non-zero change that could realistically happen in 1 year? ____ 5 years? ____
 - When you answered the question, did you think in terms of multiplying (e.g., a percent change, an x-fold increase, etc.) or did you think in terms of adding or subtracting (an x unit change)?
3. Do you expect annual differences (additive) to become larger as the magnitude of the measure increases? Exponential? If yes, this suggests multiplicative is more appropriate.
4. Do you expect non-zero changes to be constant over time? (a set amount added or subtracted each year throughout the monitored time)
 - a. Do you expect a simple linear regression model to be appropriate over the span of monitoring?
 - b. Do you expect a simple linear regression model to be appropriate during some sub-intervals?

Note: The web applet presented in Appendix A.2 may be helpful in visualizing different options to answer these questions.

What qualifies as meaningful change in the quantitative measure?

1. Assuming you could know the true value of the measure over the wilderness area (i.e., ignoring measurement error, sampling error, etc.):
 - a. What is the *smallest* annual change (in absolute value) that would be considered meaningful change in terms of wilderness character, based on professional expertise?

- b. What is the *largest* annual change (in absolute value) that would be considered to *not* be meaningful in terms of wilderness character, based on professional expertise?
- c. When you answer the previous two questions, are you considering yearly changes in raw data values for this? Or averages? Or an estimated regression slope? Or other?
- d. Do you expect year-to-year variability in the value of the measure (not due to measurement or sampling error) and that you don't think should count toward assessing a trend?

Note: The answers to these questions can be used with the web applet presented in Appendix A.3 to construct a backdrop conveying regions of meaningful change.

How are/were data collected?

1. How are/were data collected that inform the measure?
 - a. What is the primary goal the data are/were collected to support?
 - i. What are the differences between the primary goal and the WCM goal?
 - ii. What are some limitations with the data collected relative to the WCM goal?
 - b. How is the sampling performed over space?
 - i. How was the sampling design developed?
 - ii. Is a probability-based sampling design employed? Convenience sampling? Some combination?
 - iii. What are potential limitations of the sampling plan over space relative to the WCM goal?
 - What is the extent of sampling relative to the whole area of interest? And, how does the area sampled potentially differ from unsampled areas?
 - Do areas sampled change over time? (see next questions)
 - iv. How much variability is expected over space (for a fixed time interval)?
 - Assuming nothing changes in the underlying system, how much change might you still see in the observed measure between two spatial sampling units (just due to "noise" in the data over space)?
 - c. How is the sampling performed over time (frequency of data collection)?
 - i. How often is sampling performed?
 - ii. How much variability is expected over time (for a fixed spatial sampling unit)?
 - Assuming nothing changes in the underlying system, how much change might you still see in the observed measure between two time points (just due to "noise" over time (i.e., measurement error and other sources of background variability)?
 - d. How is sampling over time related to sampling over space?
 - i. Is it possible because of the design that changes observed over time are confounded with changes observed over space?
 - ii. What are the trade-offs between holding the spatial area sampled constant over time vs. sampling greater total area by moving to different areas over time?
 - iii. What are the tradeoffs between sampling a smaller area more often, or a larger area less often?
 - iv. How large is background variability in the measure over spatial units compared to background variability over chosen time intervals?
2. How large do you think the combination of measurement error and/or sampling error might be relative to your answers in the previous section (what qualifies as meaningful change)?
 - Are yearly data values used directly as the measure, or is a summary created first (e.g., 3 year average)? For example, if a summary such as an average is to be used, return to the above question

and answer it using knowledge of the summary measure (averages are less variable than raw data values).

- Note: It is possible that for the purpose of WCM less precision in quantitative estimates is needed than for other objectives (e.g., Inventory and Monitoring Program estimates of trend). The goal is to define and then pick up when enough change happens to qualify as a shift in wilderness character (which is inherently subjective and qualitative). That is, the goal is not to precisely measure any change, but to capture when enough change has occurred that wilderness character is affected or thought to be soon affected.

What are the properties of the measure?

3. Is it categorical (measure value is one of a set of categories)?
 - a. Are category labels numbers? (e.g., integers)
 - i. How many unique categories?
 - ii. What are the integer labels?
 - iii. Is zero included?
 - If yes, does zero cover only one category, or multiple categories (e.g., none *and* unknown). Note: If zero covers both non and unknown think through the potential implications of using 0 for unknown in terms of decisions.
 - b. Are category labels descriptions and not numbers?
 - i. How many categories?
 - ii. Are the categories ordered?
 - iii. Are the categories created to describe current trend in the WCM context? (e.g., categories representing the overall trend in total infested area as included as part of the Index of non-native plant species described in the *Draft Technical Guide* for Plants – pg 43)

The categorical measure described in the *Threshold for Change* section of the *Draft Technical Guide* is an example of this type of categorical measure. There are four ordered categories defined by “none”, low-, moderate-, or high-estimated percentage of wilderness occupied by nonindigenous plants.” Therefore, the categories are based on a quantitative assessment of a continuous proportion of area (Question 5 a), but the measure ultimately selected is one step removed from the quantitative measure and does not require knowing the estimated proportion, just a professional judgment of where the proportion falls in relation to the categories. It is assumed all categories are possible theoretically, but still important to gauge which categories may be realistically recorded in the near and long term. For example, for GAAR, it may be many years before even the shift from ‘none’ to ‘low-’ would be expected even if multiple non-native species were identified.
4. Is the measure a count? (e.g., # of non-native plant species)
 - a. What is the range of realistic values it can take on?
 - b. Is zero included?
5. Is it an index that takes only integer values? (e.g., a non-native plant index)
 - a. What are the unique values it can take on? What are the theoretical minimum and maximum?
 - b. What are the realistic practical minimum and maximum for the measure for the wilderness area and time period in question?
6. Is the measure defined by a continuous scale? (e.g., area, distance, etc.)
 - a. Does it only take on non-negative values? (e.g., area, distance)
 - b. What are the theoretical minimum and maximum?

- c. What are realistic practical minimum and maximum for the measure for the wilderness area and time period in question?
- d. Might change in the mean of the measure be approximately linear over time on the original scale? Linear on the log scale?
- 7. Is it a proportion?
 - a. Is it a continuous proportion? (e.g., proportion area covered) [Note- this is equivalent to a continuous percent if multiplied by 100]. “Continuous” proportions can theoretically take on any value between 0 and 1 (or 0 and 100 for percents) and are created by a ratio of two continuous measurements (like area).
 - i. What is the range of realistic values for the measure?
 - ii. Do the realistic min and max get close to 0 or 1? Depending on how the measure is constructed, is it possible to get values less than 0 or greater than 1?
 - b. Is it a proportion (or percent) that is constructed as a ratio of counts? That is, is it a count of “successes” out of a total number of “trials”?
 - i. Are the numerator and denominator recorded? (is the number of successes known? And the total number of “trials”?)
 - ii. Is zero a possibility for the numerator? (i.e., no successes and proportion of 0)
 - iii. What are the ranges of realistic values for the numerator (# of “successes” and the total number of “trials”?). What are realistic values for the proportion obtained from these?
- 8. Is the measure an average over multiple data values?
 - a. Is it an average of values collected over space?
 - b. Is it an average of values collected over time?
 - i. Is it an average of values over time within a year or is it an average over multiple years? (e.g., is the plan to use a 3-year moving average?)
 - ii. If it is over multiple years, how many?
- 9. Is the measure a quantity estimated from a statistical model?
 - a. If so, what model and parameter were used? For example, if a linear regression model is used, one could think of the measure as the estimated slope parameter since and tie the threshold for change to that parameter. As is currently explained in the guidance, this subtle difference is not recognized.
- 10. How often are measures to be input in the WCMD? If collected every year and reported every 5 years, will a summary measure be reported?

Baseline measure value

- 1. What is the current baseline measure value?
- 2. Do the properties of the baseline measure value match those of future values that will be input into the WCMD (e.g., a single data value vs. an average over multiple years or a value from a small area compared to a larger area)?
 - a. If not, what are the differences?
 - b. Is there a way to reconcile the differences or justify that the comparison is still reasonable? For example, if averages are going to be used (e.g., 3-year moving averages), this information should be explicitly included in the WCMP and a threshold for change should be developed according to the average.

A.2 Percent change and additive change

Objective: A web applet provides a quick way to visualize a user provided additive change and a percent (multiplicative) change over time, either from a baseline measure value or from successive values over time. It also provides the option to round to integers to give a more realistic perspective for count data (e.g., number of species).

Goal: This app is meant to help users quickly visualize different additive and percent changes for a given measure, with the goal of aiding in the process of deciding what should qualify as meaningful change, and to assess if a percent change approach is reasonable for the measure.

Link to applet: <https://critical-inference.shinyapps.io/PercentChangeVizApp/>

Inputs: Number of observations over time (n), reporting interval (years), baseline measure value, percent change threshold, indicator (Yes/No) of whether measure is integers only.

Outputs: Plot of percent change over time, assuming the same percent change between each reported value. If just compared to the baseline, specify just 2 values and the appropriate length of time between the baseline measure value and the value as a percent change.

Other things to note:

- The default trajectory is not realistic when the measure is a count (must be an integer) and therefore the option is provided to show the results rounded to the nearest integer (last input).

A.3 Meaningful change backdrop

Goal: Provide an accessible way to create a picture conveying regions of meaningful change before analysis, as well as the ability to display intervals of statistical uncertainty after analysis. See Section 2.5.4.4. for an example of its use within the body of the report.

Link to the applet:

<https://critical-inference.shinyapps.io/MeaningfulChangeBackdrop/>

Inputs: Minimum and maximum values to cover the scale of interest for the measure in question, the smallest value definitely considered to represent meaningful change and the largest value definitely considered to represent changes not considered meaningful. If change of interest is negative, the absolute value is considered in the same way.

Outputs: The backdrop picture showing the region of values considered meaningful change, those not considered meaningful change, and the gray area transition between.

Other things to note:

- The app does not currently allow for situations where meaningful change occurs in both positive and negative directions. This can be added if requested.

A.4 Translate p-value threshold

Objective: As discussed in the body of the report (mainly Sections 2.3.3.3, 2.3.4, 2.5 and 2.6), using “regression and statistical significance” to define a *threshold for change* results in a lack of transparency regarding what threshold is employed on the scale of the measure, making it difficult to justify and difficult to critically evaluate. To help interested readers translate a p-value based threshold into a threshold defined by the regression slope parameter (a scale that is more readily interpretable), a web applet is provided. The applet translates the model and p-value inputs into a magnitude of the regression slope parameter (under stated assumptions) and plots the associated slope under different levels of desired width of confidence intervals (i.e., precision).

Goal for its use: The web applet is meant as an investigative tool to improve transparency of the p-value threshold method. It is *not* meant to encourage continued use of the approach. In fact, it is hoped that the app will lead to an increased understanding of the limitations of the approach and be used as justification to abandon it in favor of a more context-dependent approach, as discussed in Section 2.5.4.4 and supported by the web applet presented in Appendix A.3.

Link to the applet:

https://critical-inference.shinyapps.io/Translate_threshold/

Inputs: Number of values (n), reporting interval (years), residual standard deviation (SD), alpha-level, y-intercept of lines for plotting slopes, y-axis limits (min and max) for plotting, and whether one random realization of data from the model should be plotted along with the lines.

Outputs:

- Graph of the resulting slopes from higher desired precision (red) and lower desired precision (orange)
- A random realization to accompany the higher desired precision line (if user opts for this)
- Translations and interpretations for both higher and lower precisions in text below the plots
- A statement about assumptions

Other things to note:

The mathematical justification for the translation is provided on the following two pages.

Documentation: Translating of alpha-level threshold to slope parameter threshold

This documentation accompanies the web applet developed to convey translation from a threshold based on comparing a two-sided p-value to an alpha level to a threshold based on magnitude of the slope parameter of simple linear regression (SLR) model. The p-value in question is obtained from testing the null hypothesis that the slope parameter is equal to zero (and under all assumptions of the SLR model).

The web applet can be found here: https://critical-inference.shinyapps.io/Translate_threshold/

Overall strategy

The translation requires (1) assuming the SLR model is correct, (2) specifying a residual standard deviation (background variability), and (3) specifying a desired precision for the estimate. The third part would often be done via specifying a type II error rate or power, but I prefer the approach of using confidence interval width or half-width, particularly in the context of specifying meaningful change. See Chapter 23 in *The Statistical Sleuth: A Course in Methods of Data Analysis* (2013, 3rd Edition, Brooks/Cole Cengage) by Ramsey and Schafer for more detail on this choice and comparisons to power.

The general strategy is to set a width or half-width desired and then solve, using the known sample size, a desired confidence level, and the components that go into calculating the standard error of an estimated SLR slope coefficient to solve for the smallest (in absolute value) value that is assumed to be practically meaningful.

Sample size formula

The approach can be thought of as based on the sample size formula associated with simple linear regression (see Page 71 in *The Statistical Sleuth* for more information.) The formula is based on specifying a desired width (or half-width) of a $(1 - \frac{\alpha}{2})\%$ confidence interval associated with the slope parameter. It combines the formula for a confidence interval and the formula for the standard error of the estimated slope in SLR.

$$n = \frac{4[t_{df}(1 - \frac{\alpha}{2})]^2 s_r^2}{\delta^2 s_x^2} + 1$$

Where δ is the width of the $(1 - \alpha)\%$ confidence interval, s_r is the residual standard deviation, s_x is the standard deviation of the x variable (years in this case), and $t_{df}(1 - \frac{\alpha}{2})$ is the $(1 - \frac{\alpha}{2})$ quantile of a t distribution with the specified $df = n - 2$.

The above formula is then solved for δ , which provides the confidence interval width that would be associated with given values for all the other inputs (n , α , s_x , s_r , and t_{df}):

$$\delta = \sqrt{\frac{4[t_{df}(1 - \frac{\alpha}{2})]^2 s_r^2}{n s_x^2} + 1}$$

And, the half-width is simply half of δ , or

$$\delta_{half} = \sqrt{\frac{[t_{df}(1 - \frac{\alpha}{2})]^2 s_r^2}{n s_x^2} + 1}$$

Interpretation

The confidence interval width solved for (δ) provides the smallest value of a slope such that, for the given inputs and assumptions, the confidence interval cannot contain both 0 and δ . In other words, δ can be interpreted as the assumed practically meaningful change that comes from comparing the p-value to the stated α -level. The halfwidth corresponds to the same thing, but with less desired precision – the constraint is only that the halfwidth can't cover both the assumed practically meaningful value and zero. See more in Chapter 23 of *The Statistical Sleuth* about the justification and philosophy behind this approach. The half-width approach is generally closer to what would be obtained with power levels (desired Type II error rate) around 0.5 and using the whole width corresponds power of around 0.90 (type II error rate of 0.1).

The R function written to calculate this for the web applet is provided here

```
SLR_threshold_CIwidth <- function(n, t_interval, sd_resid, alpha) {
  x_vec <- seq(0, ((n-1)*t_interval), by=t_interval)
  sd_x <- sd(x_vec)
  w <- sqrt(((4*qt((1-(alpha/2)), df=n-2)*(sd_resid^2))/(n*(sd_x^2))) + 1)
  #hw <- w/2
  return(w)
}
```

APPENDIX B – Original Scope of Work (SOW)

The work will address two general objectives: (1) review the *Threshold for Change* and *Determining significant change and trend in a measure* parts of the WCM process and provide additional or alternate recommendations, and (2) use 2-3 proposed GAAR measures to illustrate potential cases and demonstrate implementation of recommendations. It is important to consider the *threshold for change* part of the process early in the evaluation of the usefulness and relevance of proposed measures (i.e., it should be carried out before measures are decided upon).

The following tasks are defined to address the objectives within the constraints of the available budget.

Task 1: Work with GAAR to identify 2-3 proposed measures that will provide useful contexts to discuss challenges and limitations of the WCM process and that represent challenges that will arise with other measures and in other parks, and that help GAAR move forward in justifying and finalizing their measures.

Update: *In the course of the work, the scope of this task evolved so that the review is now more general and meant to provide useful information (in Task 3) that applies to more than 2-3 measures. A list of current thresholds for change for all measures provided in the current GAAR draft are included in Section 1.3. After starting a more in-depth review of the information available, the discussion and recommendation of the review became more general and applicable across many measures and parks. Because of this, less information is provided about specific measures. This was agreed upon with Kristin Pace and Adrienne Lindholm*

Task 2: Review the *Threshold for Change* part of the currently recommended WCM process, as described in *Keep It Wild 2, Draft NPS Wilderness Character Monitoring Technical Guide* (see pages 19-22), and the current Draft *Gates of the Arctic Wilderness Character Baseline Assessment and Monitoring Plan*. Write a general review of the recommendations from a statistical and decision-making perspective, highlighting positive aspects as well as limitations and challenges in the current guidance.

Task 3: Develop written recommendations to improve the defining meaningful change and threshold part of the process. Use case studies and a tutorial approach to demonstrate how to carry out the recommendations using the 2-3 selected GAAR measures. If deemed useful and resources allow, develop simple web applets to help users carry out the recommended process.

Task 4: Email correspondence and meetings associated with the work

Task 5: Revisions of draft deliverables, presentation of the work, training, etc.