NIWA
Taihoro Nukurangi

Attribute states and uncertainty

Preliminary expert commentary on implementation of clause 3.10(4) of the NPS-FM 2020

Prepared for MBIE Envirolink and Horizons Regional Council

June 2023

Climate, Freshwater & Ocean Science

ISBN: 9781991061232

Report number: 2023/EXT/1804

Prepared by:
Juliet Milne, Ned Norton (LWP), Doug Booker, Paul Franklin, David Wood, Olivier Ausseil (Aquanet Consulting), Roger Young (Cawthron Institute), Michael Patterson (Horizons Regional Council) and Luke Fullard (Horizons Regional Council)

For any information regarding this report please contact:
Juliet Milne
Resource Management Scientist
+64 4 386 0901
juliet.milne@niwa.co.nz

National Institute of Water & Atmospheric Research Ltd
Private Bag 14901
Kilbirnie
Wellington 6241

Phone +64 4 386 0300

NIWA CLIENT REPORT No: 2022364WN
Report date: June 2023
NIWA Project: ELF23301

Revision	Description	Date
Version 0.1	Draft in preparation/in review	14 February 2023
Version 0.2	Revised draft version sent to client	18 April 2023
Version 1.0	Final version sent to client	30 June 2023

Quality Assurance Statement		
	Reviewed by:	Dr Scott Larned
	Formatting checked by:	Juliet Milne
	Approved for release by:	Dr Judy Sutherland

Contents

Tables

Figures

Executive summary

A key step in the process of implementing the National Policy Statement for Freshwater Management 2020 (NPS-FM) is *"to identify the baseline state"* for a suite of attributes (clause 3.10). Attributes are measurable characteristics (e.g., nitrate-nitrogen concentration) used to assess the extent to which a certain freshwater value (e.g., ecosystem health) is provided for within a river or lake. Under the National Objectives Framework (NOF) of the NPS-FM, attribute state is generally a summary statistic (e.g., median) that is calculated from measurements or observations made regularly (e.g., monthly) over a fixed period (e.g., five-years). The calculated attribute state can also be expressed as a category that represents one of four or five numerically defined bands (A to D or A to E). In many cases, the NOF stipulates a minimum acceptable state (or national bottom line) for an attribute.

Baseline attribute states (BAS) provide benchmarks against which regional councils and communities can set the future target states that attributes must attain to achieve the environmental outcomes councils have set for freshwater values. The setting of a target attribute state (TAS) must be consistent with the direction of Policy 5 of the NPS-FM to at least *maintain* the health and well-being of water bodies and, where they are degraded, *improve* them. This means that, at a minimum, a TAS cannot be set lower than a BAS. Further, where a BAS is worse than the national bottom line (NBL) specified for an attribute in the NPS-FM, the TAS must be set above the NBL. Councils must monitor and regularly assess current attribute states (CAS) to track progress towards meeting the corresponding target states. The datasets used to estimate baseline and current attribute states for a given site take the form of time-series, or repeated observation of an attribute value made at multiple points in time over a given period (e.g., monthly observations at a river monitoring site over a five-year period).

Subclause (4) of clause 3.10 of the NPS-FM states:

> *"attribute states and baseline states may be expressed in a way that accounts for natural variability and sampling error".*

The terms "natural variability" and "sampling error" are not defined and existing national guidance does not address how subclause (4) might be implemented by councils. Horizons Regional Council, on behalf of all regional councils, sought an MBIE Envirolink advice grant for NIWA to convene an expert panel to discuss how clause 3.10(4) might be implemented. The primary tasks of the expert panel were to:

- define "natural variability and sampling error" in the context of NPS-FM attribute state assessments, and
- determine how numeric attribute state might be expressed in a way that accounts for natural variability and sampling error.

It was acknowledged in the advice grant application that the workshop discussions would likely only lead to preliminary advice to support implementation of clause 3.10(4). As such this report represents a starting point to inform the development of any subsequent national guidance.

Clause 3.10(4) – intent and key terms

Clause 3.10(4) clearly applies to baseline attribute states. The term "attribute states" in subclause (4) is not defined and therefore could also relate to current attribute states and/or target attribute states).

Natural variability is the variation in values of an attribute caused by natural processes. For example, river flows and nutrient concentrations exhibit natural variability over space and time due to weather, climate, and physical and ecological processes. Natural variability is independent of human influences or measurement error. The focus of the expert panel for the purpose of this report was primarily on temporal variability in attribute states at the scale of individual monitoring sites.

Natural variability of an attribute occurs over a range of timescales (e.g., diel cycles, seasonal cycles, and interannual variability associated with naturally occurring trends or climatic cycles). In addition, natural variability is associated with irregularly spaced events such as floods or droughts (where these are not anthropogenically-driven). Variability at interannual time-scales can be seen in attribute time-series plots, such as those presented in this report, often in the form of monotonic trends and cyclic fluctuations.

The NPS-FM requires a BAS to be set with a specific end date, but no starting dates or assessment periods are specified. Natural drivers can generate cyclic fluctuations in attribute state, and the period used for estimating a BAS can coincide with different portions of the cycle (e.g., at or near a peak or trough). The estimated BAS has important consequences for councils and communities given that the corresponding TAS must be set at or better than the baseline state, and this in turn will influence the nature and extent of limits imposed on resource use.

Not all temporal variability is natural, and natural and anthropogenic drivers can interact, which makes partitioning natural and anthropogenic variability impractical. For this reason, the term “environmental variability” might be more appropriate than “natural variability”.

Sampling error is statistically defined as the difference between a sample statistic used to estimate a population parameter and the actual, but unknown, value of that parameter. In the context of the NPS-FM, the sample statistic is the NOF summary statistic (e.g., median, 95th percentile) used to represent the numeric attribute state. This statistic is only an estimate of the ‘true’ attribute state because it is calculated from a limited number of measurements over a finite assessment period. This means that uncertainty due to sampling error will always be associated with sample-based estimates of attribute state.

The uncertainty associated with sampling error is commonly quantified using a measure of precision, such as a confidence interval (CI) around the estimated attribute state. The use of CIs and other inferential statistics requires that the samples comprising the time series are independent and that the distribution is stationary (e.g., the statistical properties of the time series do not change over the assessment period). In the case of NOF attribute time-series, these requirements are likely to be violated, as indicated by long-term trends and seasonal and inter-annual fluctuations.

Implementing clause 3.10(3) and 3.10(4) – interim advice

The word “identify” in clause 3.10(3) provides flexibility in how a BAS is estimated and the process of estimating BAS will need to involve elements of expert judgement. We assume that, where possible, a BAS should be calculated as a summary statistic from existing monitoring data over a finite time period. As a rule-of-thumb, a five-year period will provide a robust estimate of attribute state where sampling is monthly. Calculating BAS over a longer period of time may be appropriate if a council has a consistent monitoring record and is confident that temporal changes in attribute state are solely attributable to natural drivers. In practice it may be difficult to establish this, and using a BAS assessment period longer than five years increases the likelihood that the BAS estimate is influenced by anthropogenically-driven changes in state.

It is optional under clause 3.10(4) to account for natural variability and sampling error when expressing attribute states. While this means that a council could decide to express BAS as a single number (i.e., the NOF summary statistic only), there may be substantial uncertainty associated with estimates of attribute state. Because interpretations of, and comparisons between, baseline, current and target attribute states are influential in council freshwater management decision-making, scientists should acknowledge and communicate uncertainty associated with attribute state. Whether the uncertainty is expressed alongside a single numeric value in a regional plan or in a background supporting technical document is a decision for council planners to make, but this information should be documented and made publicly available.

Reporting estimates of uncertainty around numeric BAS and CAS estimates is also important for subsequent NOF steps – in particular, clause 3.11(2) (setting the TAS "at or above" the BAS), and clauses 3.18 to 3.20 (monitoring and reporting progress on 'maintain or improve', and taking action where degradation in a CAS is identified). In contrast, we recommend treating the TAS as a single numerical value (i.e., without an an expression of uncertainty).

As an interim approach to acknowledging that natural variability and sampling error contribute to uncertainty in attribute state estimates, we suggest that councils provide at least a narrative description of their confidence in BAS and CAS estimates . An alternative method is to identify and assess temporal changes in the attribute state time-series with a rolling time window. If confidence intervals are used to provide numeric estimates of uncertainty, the time series data used to calculate them need to be checked for violations of the statistical requirements noted above, and pre-processed as necessary to remove trends and fluctuations, and checked for serial autocorrelation before calculating CIs.

The time period over which to evaluate uncertainty in a BAS estimate is best left to the discretion of councils because we cannot recommend one time period that will be appropriate in all cases. As noted above, the NPS-FM only provides end dates for baseline assessment periods, not start dates. Many councils have long time-series for some attributes and these time-series should be assessed to identify trends or long-term cyclic fluctuations. Examining the long-term data at the outset of BAS establishment also provides an opportunity for councils to document where the estimated BAS sits within a trend trajectory and/or cycle. This provides important context for future evaluations of a CAS against the corresponding TAS and may assist with informing management intervention or re-evaluations of the TAS.

Given the lack of readily available methods to make statistically robust comparisons of current and target attribute states, we recommend that temporal trend assessments, for which standard procedures are already well established, are used as the primary means to indicate if a CAS is on the right trajectory to meet the corresponding TAS. Both temporal trend direction and magnitude (and the associated confidence in these estimates) are important considerations.

Information that should be provided to decision makers to help them to interpret the results of attribute state and trend assessments includes:

- whether there has been a change in catchment land use and/or management,
- whether a change in attribute state or trend is also evident in another attribute that it may influence, or at unimpacted/reference sites within the same area/catchment,
- whether long-term climate cycles or extreme events may have influenced attribute state over the assessment period, and

- whether there have been any changes in sampling and/or measurement methods that might have impacted the attribute assessment.

Ultimately, expert judgement will be needed to evaluate and interpret changes in attribute state.

Next steps

Formal national guidance needs to be developed to support expressions of freshwater attribute state *"in a way that accounts for natural environmental variability and sampling error"*. In addition, more work is needed to understand, quantify and account for the influence of cyclical climate processes and other drivers of the natural variability associated with attribute state (and trends) through time. Three pieces of work that may assist councils to better account for environmental variability and sampling error are:

- investigating statistical methods to remove the effect of long-term trends and seasonal and interannual fluctuations in attribute time-series data to evaluate the residual variation about the estimated attribute state,
- investigating new proxy measures of climate variability that may be correlated with variation in attribute state, and
- investigating methods to characterise variation in time-series data from non-stationary distributions.

This report has focussed on temporal variability at the scale of a single site but spatial variability forms a significant component of natural variability. National guidance is also needed to assist councils with accounting for and expressing spatial variability in the context of clause 3.10(4).

Given the complexities involved with accounting for natural variability and sampling error when establishing and comparing attribute states, some details of attribute state assessments and reporting specified in the NPS-FM 2020 may need to be revisited.

1 Introduction

A key step in the process of implementing the National Policy Statement for Freshwater Management 2020 (NPS-FM) is determining the baseline state for (at a minimum) a suite of 22 mandatory water quality (e.g., *E. coli*) and biological (e.g., periphyton biomass) attributes. Clause 1.4 - Interpretation, of the NPS-FM defines baseline attribute state as:

> *"the best state out of the following:*
>
> *(a) the state of the attribute on the date it is first identified by a regional council under clause 3.10(1)(b) or (c)*
>
> *(b) the state* of the attribute *on the date on which a regional council set a freshwater objective for the attribute under the NPS-FM 2014 (as amended in 2017)*
>
> *(c) the state of the attribute on 7 September 2017."*[1]

The above dates represent end dates for baseline assessment periods. No start dates for assessment periods are specified. Consequently the length of baseline assessment periods are also unspecified.

Subclause (4) of clause 3.10 of the NPS-FM states that "*attribute states and baseline states **may** be expressed in a way that accounts for natural variability and sampling error".* The inclusion of this clause recognises that an attribute state is estimated from a limited number of measurements (observations) made over a fixed time period, and is subject to variability in time and space. Various factors (e.g., sampling and measurement technique, flow, climate) contribute to variability in attribute measurements, some of which are random (stochastic) and, therefore, unpredictable. It is important that these sources of variation or inherent 'noise' are considered alongside the numeric attribute state when setting the future desired state (target attribute state). Policy 5 of the NPS-FM directs that, at a minimum, a target attribute state (TAS) cannot be set lower than the baseline attribute state (BAS). Further, where a BAS is worse than the national bottom line specified for an attribute in the NPS-FM, the corresponding TAS must be set above this. Councils must monitor and regularly assess current attribute state (CAS) to track progress towards meeting the corresponding TAS.

Horizons Regional Council, on behalf of all regional and unitary councils (hereafter *regional councils* or *councils*), sought an MBIE Envirolink advice grant (HZLC166) for NIWA to convene an expert panel workshop on how to interpret and implement clause 3.10(4). The primary tasks of the expert panel were to:

1. Define "natural variability and sampling error" in the context of NPS-FM attribute state assessments.
2. Determine how numeric attribute states might be expressed in a way that accounts for natural variability and sampling error, including:
    - whether this is on an attribute-by-attribute and site-by-site basis or by, for example, some grouping of attribute and site types, and
    - the most appropriate statistical expressions to use (e.g., confidence intervals, standard deviation, coefficient of variation).

It was also envisaged that workshop discussions might address what, if any, consideration should or can be given to the impacts of unusual/extreme hydrological conditions (e.g., droughts, floods) and

[1] This text incorporates minor amendments to the NPS-FM that were released in December 2022 and came into effect on 5 January 2023. The amendments made to clauses 1.4 and 3.10 of the NPS-FM have no material consequence to the commentary provided in this report.

natural climate cycles (and potentially climate change) on variability in attribute state. It was acknowledged in the advice grant application that the workshop discussions would be preliminary only and that further work would be needed to develop and refine any implementation approaches identified by the expert panel.

This report provides a summary of the key material discussed by the expert panel and sets out preliminary commentary to support councils in implementing clause 3.10(4) of the NPS-FM.

1.1 Approach

The following group of experts spanning multiple science disciplines and organisations attended an on-line workshop on 3 November 2022:

- Dr Doug Booker (NIWA, Hydro-ecological Modeller)
- Dr David Wood (NIWA, Water Quality Scientist)
- Dr Paul Franklin (NIWA, Freshwater Ecologist)
- Juliet Milne (NIWA, Regional Management Scientist)
- Dr Ton Snelder (LWP, Scientist)
- Ned Norton (LWP, Water Resource Management Consultant)
- Dr Roger Young (Cawthron, Manager Freshwater Ecosystems)
- Dr Olivier Ausseil (Aquanet Ltd, Principal Scientist).

Senior science staff from Horizons Regional Council (Mike Patterson, Maree Patterson and Dr Luke Fullard), Bay of Plenty Regional Council (James Dare), Environment Canterbury (Shirley Hayward) and Auckland Council (Dr Coral Grant) also attended the workshop. These council staff have been considering clause 3.10 implementation and some had examples of baseline attribute state assessments and issues to contribute to the discussion.

Background material was pre-circulated to all workshop participants to establish a starting point for workshop discussions. This material was prepared by Juliet Milne, with input from Mike Patterson, Ton Snelder, Ned Norton and Doug Booker. The material was considered a working draft and included:

- an overview of clause 3.10 and a discussion of its likely intent,
- suggested definitions for natural variability, sampling error and other key terms,
- examples from Horizons Regional Council as to the issues with application of clause 3.10(4),
- some suggested approaches Horizons Regional Council and others have considered in expressing the variability associated with attribute state metrics, and
- a series of questions for workshop participants to consider prior to the workshop.

For manageability, workshop discussions were focussed primarily at the scale of single sites at which monitoring data are collected over time, and spatial variability and modelled estimates of attribute state were not considered.

Following the workshop, Juliet Milne prepared a written summary of the workshop that was circulated for comment by attendees. A subgroup comprising Mike Patterson, Ned Norton and Juliet Milne then developed a draft strawman for communicating estimates of attribute state and expressing the uncertainty associated with these estimates. The draft strawman was circulated to the wider panel ahead of a follow-up on-line discussion on 6 December 2022.

This report consolidates the preliminary background material, the workshop and strawman discussions, and some subsequent thinking. The examples provided to illustrate how different attribute states and associated uncertainty might be expressed will need testing by regional council practitioners and possible modification to ensure that they are fit for purpose.

1.2 Terminology

The reader is referred to the glossary for definitions of some specific terms used in this report.

1.3 Report outline

This report comprises three further sections. In section 2 we examine clause 3.10 of the NPS-FM in detail, including how it relates to different types of attribute states (BAS, CAS and TAS), and what is meant by “natural variability” and “sampling error”. In section 3 we provide preliminary commentary to support implementation of clause 3.10(3) and 3.10(4) and present examples for a possible interim approach to (i) estimating a BAS and characterising the associated uncertainty, (ii) setting a TAS, and (iii) estimating a CAS and characterising the associated uncertainty, and (iv) determining whether a CAS is on track to meet the corresponding TAS. We also outline some additional considerations for attribute state assessments. In section 4, we present some brief conclusions and recommendations for further work.

2 Clause 3.10 – intent and key terms

The NPS-FM 2020 was introduced as part of the Government's *Essential Freshwater* package. One of the three primary aims of this package, "*stop further degradation and loss*", provides important context for identifying baseline attribute state under clause 3.10. This clause sits within the wider National Objectives Framework (NOF) process (Figure 2-1) that directs how regional councils, with communities and tangata whenua, are to manage freshwater in their regions. How clause 3.10 is implemented has implications for subsequent steps in the NOF process. In this section we examine clause 3.10, outline the role of the three different types of attribute state, and discuss our understanding of "natural variability" and "sampling error."

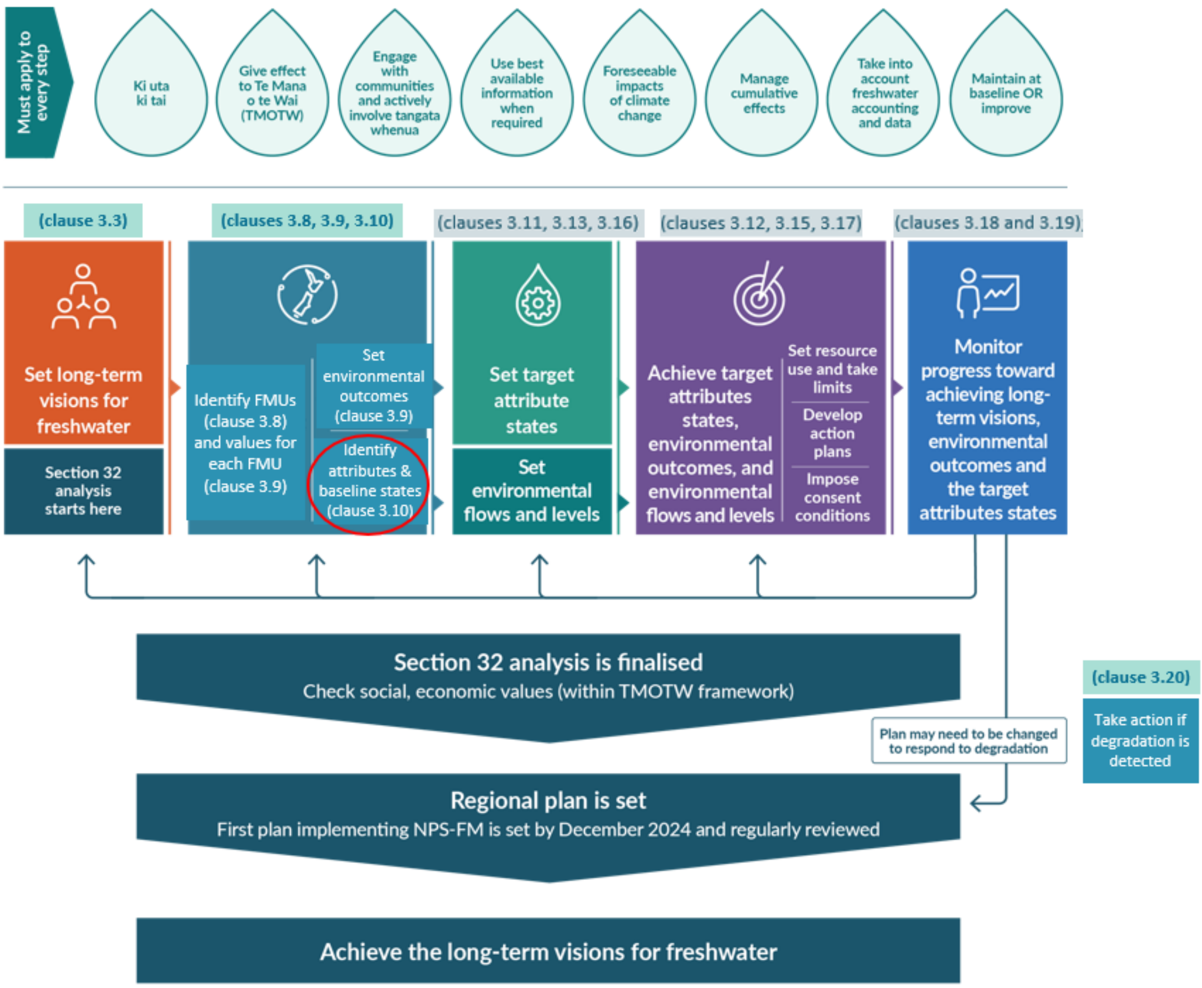


Figure 2-1: Where clause 3.10 (indicated by the red circle) sits within the wider NPS-FM NOF process. Adapted from MfE (2022).

2.1 What clause 3.10 says

Clause 3.10, reproduced in Figure 2-2, has four subclauses that can be separated into two steps:

- identification of attributes (subclauses 1 and 2), and
- identification of attribute baseline state (subclauses 3 and 4).

3.10 Identifying attributes and their baseline states, or other criteria for assessing achievement of environmental outcomes

(1) For each value that applies to an FMU or part of an FMU, the regional council:

(a) must use all the relevant attributes identified in Appendix 2A and 2B for the compulsory values listed (except where specifically provided otherwise); and

(b) may identify other attributes for any compulsory value; and

(c) must identify, where practicable, attributes for all other applicable values; and

(d) if attributes cannot be identified for a value, or if attributes are insufficient to assess a value, must identify alternative criteria to assess whether the environmental outcome of the value is being achieved.

2) Any attribute identified by a regional council under subclause (1)(b) or (c) must be specific and, where practicable, be able to be assessed in numeric terms.

(3) Every regional council must identify the baseline state of each attribute.

(4) Attribute states and baseline states may be expressed in a way that accounts for natural variability and sampling error.

Figure 2-2: Clause 3.10 of the NPS-FM 2020, as amended in December 2022. Subclauses 3 and 4 (highlighted) are the focus of this report.

Under subclause (3) the baseline state of each attribute must be "identified" but there is flexibility in how this is done (reflected by the words 'other criteria' in the clause title). In accordance with clause 1.6 (and prior to the December 2022 NPS-FM amendments, stated explicitly within 3.10(3)), councils must make use of "the best available information" to identify a BAS. This recognises that few or no measured data may be available for estimating the baseline states for some attributes.[2] While MfE (2022) guidance directs councils to use real (measured) data where available, it also explicitly recognises that modelled data and other approaches will be needed. Further, clause 1.6 of the NPS-FM gives councils discretion to interpret the best information available *"in the way that will best give effect to"* the NPS-FM. In our view, that leaves flexibility in how attribute states estimated from models or short time-series are both defined and communicated.

Clause 3.10(4) clearly applies to at least BAS. The term "attribute states" in subclause (4) is not defined and could relate to current attribute state and/or target attribute state. Consequently, as written, **it is optional** under subclause (4) to account for natural variability and sampling error associated with BAS and other attribute states. We outline these different attribute states next.

2.2 Attributes and attribute states

Attributes are defined in the NPS-FM as measurable characteristics (e.g., nitrate-nitrogen concentration, periphyton biomass) used to assess the extent to which a certain freshwater value (e.g., ecosystem health) is provided for. The NPS-FM refers to three forms of attribute state: baseline attribute state (BAS), current attribute state (CAS) and target attribute state (TAS). Only BAS is defined in the NPS-FM. As noted in Section 1, end dates for BAS assessment periods are set out in

One panel member suggested that the statement could be interpreted in a way that means a council could justify using a long time-series if this was available and potentially more informative for estimating a BAS.

the NPS-FM, but start dates and lengths of assessment periods are not. Implications of variable assessment periods are discussed in subsection 3.2.2.

Attribute states can be expressed numerically and as categories that represent one of four or five numerically defined bands (A to D or A to E). For most NOF attributes, a minimum acceptable state (or national bottom line) has been set.

In the following text, we outline our understanding of the roles of baseline, current and target attribute states, drawing on a conceptual model provided by Horizons Regional Council.

2.2.1 Baseline attribute state

Depending on the attribute, a BAS serves as a 'floor' (Figure 2-3) under, or a 'ceiling' above, the corresponding TAS. It therefore provides a starting point or benchmark for evaluating management actions that may be needed to "stop further degradation."

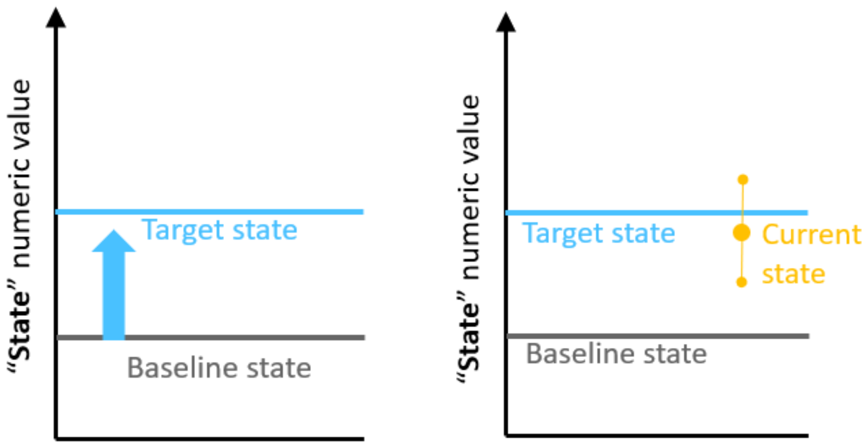


Figure 2-3: Horizons Regional Council's conceptual model of the different forms of attribute state under the NPS-FM. The model presented here applies to an attribute such as dissolved oxygen concentration where, as illustrated in the lefthand plot, BAS serves as a 'floor' and an improvement requires TAS to be set at a higher numeric value. The symbol and vertical line for CAS (righthand plot) illustrates that there is a degree of variability or uncertainty around any CAS estimate when comparing it to a proposed TAS (although not shown, there is also uncertainty associated with the initial BAS estimate).

Although the NPS-FM does not specify how to "identify" a BAS, we suggest that, where sufficient data exist, BAS should be calculated as a median (or mean) from measurements collected over a fixed assessment period. We revisit this in Section 3 and provide examples of how a BAS may be expressed where few monitoring data are available.

2.2.2 Target attribute state

Target state is the desired future state that a given attribute must attain in order to achieve the environmental outcomes councils set for each freshwater value. As directed by clause 3.11 of the NPS-FM, a TAS must be time bound[3] and set at or above the baseline state of that attribute (Figure 2-3).[4] It must also be set above a national bottom line (where one applies). Refer to clause 3.11 for the complete list of requirements for setting a TAS.

[3] Further, where the timeframe is long term (e.g., decades), councils must set interim TASs of not more than 10 years to assess progress towards achieving the (longer term) TAS.

[4] There are a few exceptions, notably *E. coli* and cyanobacteria (clause 3.11(3)) where the TAS must be set above the baseline state unless the baseline state is already in the A band.

Once a TAS has been set as part of an operative regional plan, the primary focus shifts from the starting point in terms of its corresponding BAS to measuring the current state of that attribute against its target state (i.e., to assess progress towards achieving the TAS, in accordance with subclause 3.30(2) of the NPS-FM). However, the BAS remains relevant as the initial benchmark in regional plan development against which the CAS could potentially be compared to assess whether an improvement has been made (even if the TAS has not been achieved). In addition, in some cases TAS will be set to maintain BAS (i.e., TAS = BAS).

2.2.3 Current attribute state

CAS is effectively the present (or, more correctly, recent) attribute state at any time of reporting, calculated in accordance with the summary statistic and assessment period specified in the NPS-FM (e.g., for MCI, Table 14 of the NOF specifies a median statistic based on annual sampling over 5 years). For some attributes, the assessment period is not defined in the NPS-FM.

Under clause 3.30(2)(b), councils are required to publish a comparison of the current state of attributes against the target attribute states at least every five years. However, given the requirement to publish monitoring data annually (clause 3.30(1)), some councils may intend to include comparisons of updated estimates of the CAS for each attribute against the corresponding TAS as part of their annual reporting. There is currently no national direction or guidance on how attribute state comparisons should be performed or reported. We revisit this in Section 3.

2.3 Natural variability

Although not defined in the NPS-FM, natural variability has been defined by MfE (2018) as follows:

> *"Natural variability refers to the natural variations in many aspects of the environment that we measure. For example, flows and contaminant concentrations in a river vary in time, and contaminant leaching rates vary in space. This variation is an inherent part of the environment and cannot be reduced by collecting more information…"*

There are many sources of natural variability, in addition to climate processes. The definition above indicates that there are both temporal and spatial components to natural variability in attribute states. The MfE (2022) NOF guidance acknowledges that BAS can vary across a freshwater management unit (FMU) and recommends, as best practice, that *"the baseline state should be determined as close as possible to the location where current or future monitoring sites will be located."* This statement and subsequent text recognising that different baselines states can be applied across an FMU suggests that temporal variability might be the intended focus of clause 3.10(4) (especially given subsequent NOF clauses on tracking progress through time). However, there is nothing that excludes consideration of spatial variability.

As noted in subsection 1.1, this report focuses on temporal variability at individual monitoring sites. Natural temporal variability at a site may include intra-daily variability associated with diel cycles, seasonal variability associated with seasonal cycles, inter-annual variability associated with monotonic trends and long-term cycles, and extreme variability associated with infrequent events such as floods or droughts (where these are not anthropogenically-driven).

Time-series plots provided by Horizons Regional Council science staff illustrate temporal variability in attribute state from river sites (Figures 2-4 to 2-6), including a site that is minimally impacted by anthropogenic factors (Figure 2-5). This variability, in the form of fluctuating cyclical patterns, represents non-stationarity in attribute state due to natural variability. Climate-driven changes in precipitation/flow and/or temperature are likely responsible for this variability given that the fluctuating curves in these (and many other time-series) plots correlate with 5-7 year cycles associated with the Southern Oscillation Index (SOI) (e.g., Snelder et al. 2022a, b).

A particular challenge for councils is that a BAS must be set as the best state (as defined in the NPS-FM) during a period that ends on one of three specific dates (refer Section 1) and, where monitoring data are available, is calculated over a specified time (assessment) period preceding that date. In situations where the BAS assessment period is shorter than the period of cyclical fluctuations in attribute values (often denoted by peaks and troughs, see Figure 2-5), or where attribute values are influenced by infrequent events (e.g., a period of intense floods), the BAS may not provide a representative estimate of 'average' state. This has important consequences for councils and communities given that target state must be set at or better than baseline state. For example, when considering the scale of improvement that is required to achieve a target state, councils need to consider the difference between the CAS and the TAS. If the BAS assessment period by chance was to fall into one of three periods commencing in 2003, 2004 or 2005 in Figure 2-4, then the TAS could be required to be set at or above the C, B or A band. The consequence of this could include setting overly lenient controls on resource use if TAS is set in the C band, or imposing stricter limits on resource use if TAS is set in the A band.

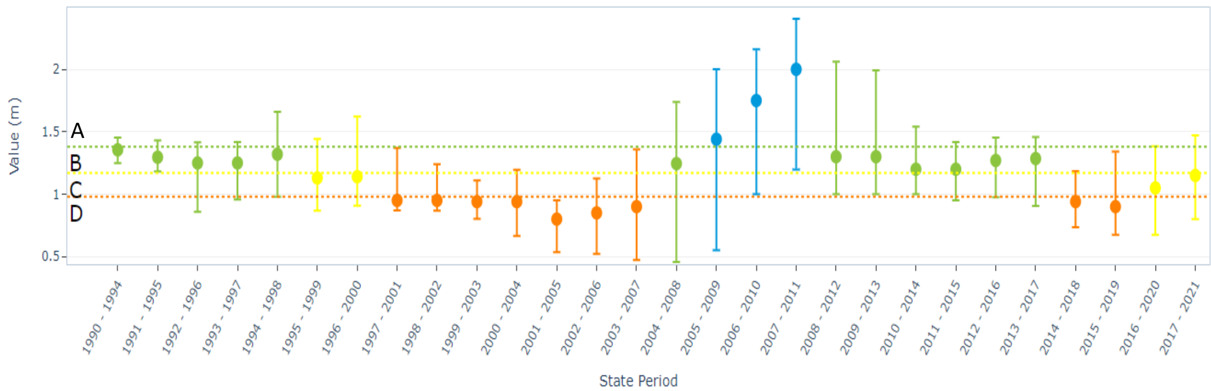


Figure 2-4: Visual clarity (median (closed circle) ± 90th percentile confidence intervals[5]) in the Manawatū River at Hopelands, based on monthly sampling and rolling five-year assessments. Colours denote the NPS-FM NOF band in which calculated attribute state sits (blue = A band, green = B band, yellow = C band, orange = D band).

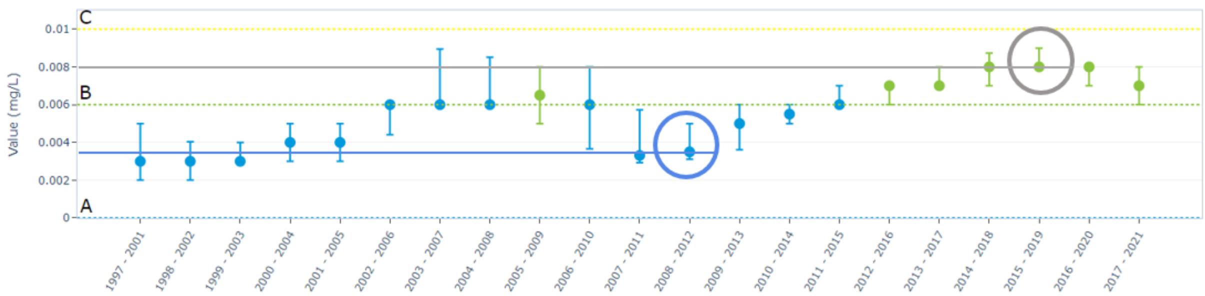


Figure 2-5: Dissolved reactive phosphorus concentrations (median (closed circle) ± 90th percentile confidence intervals[5]) for the Rangitīkei River at Pukeokahu, based on monthly sampling and rolling five-year assessments. This site is considered to be minimally impacted by anthropogenic factors such as land use. Colours denote the NPS-FM NOF band in which the calculated attribute state sits (blue = A band, green = B band). The open blue and grey circles indicate examples of a cyclic trough and peak, respectively.

[5] Note that, as discussed next in subsection 2.4, these confidence intervals do not accurately represent statistical precision, as an indicator of sampling error, because they conflate natural variability and sampling error. Some of the intervals are wide and span multiple attribute bands (especially in Figure 2-4).

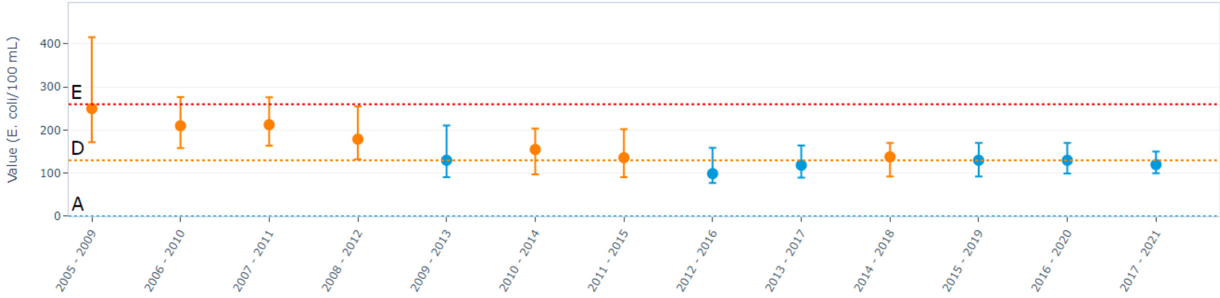


Figure 2-6: Median *E. coli* counts (median (closed circle) ± 90th percentile confidence intervals[5]) for the Manawatū River at Hopelands, based on monthly sampling and rolling five-year assessments. Colours denote the NPS-FM NOF band in which the calculated attribute state sits (blue = A band; and orange = D band).

The natural variability in median state illustrated in the time-series plots (especially Figure 2-5) can confound the detection of improvement or degradation in attribute state driven by anthropogenic factors (e.g., land use and management actions).[6] While Snelder et al. (2022b) were able to attribute temporal trends in water quality variables to the combination of an indicator of climate variability (i.e., fluctuations in the SOI) and multiple indicators of pastoral land use, different sites showed different responses to the SOI (e.g., some positive and some negative correlations) and these responses were also variable-specific. Attribution was based on statistical models that inherently involve assumptions and uncertainty.

A further complication with assessing natural variability is that not all climate variability is attributable to natural processes (because of anthropogenically-driven long-term climate change) and climate variability can also influence anthropogenic responses (e.g., an extended dry spell could lead to a change in land use or water abstraction and a subsequent increase in periphyton biomass). For this reason, the broader term of “environmental variability” might be more appropriate than “natural variability”.

Extreme events that may influence attribute state are also important to consider. For example, rare, severe droughts and floods affect attribute measurement values, and these effects may extend for long periods following the event.

It would be useful to incorporate additional proxy measures of climate variability in assessments of variation in attribute states. It would also be useful if a ‘normal’ year could be identified from a climatic point of view to contextualise assessments and reporting of attribute state. This requires further research. Until that is done, the SOI is simply used as a useful index of climate-related variability.

Natural variability also includes variability associated with seasonal cycles. Where an attribute exhibits seasonality, this will also contribute to the variability in attribute state within an assessment period. Ignoring the influence of seasonal variation when estimating uncertainty in attribute state will result in an over-estimation of this uncertainty. The influence of seasonal variability is also important when assessing changes in attribute states between assessment periods.

In summary, freshwater environments are dynamic and respond to multiple drivers. Statistical estimates of the state of freshwater attributes will, therefore, vary in time and space.

[6] Snelder et al. (2022b) developed models that related observed water quality trends (over different time durations) to climate variation (i.e., fluctuations in the SOI) and mean (and changes) of productive land use in catchments. Across 10-year windows, land use signals were generally swamped by the greater influence of climate variation.

2.4 Sampling error

Sampling error is not defined in the NPS-FM, but is a well-established statistical term. The following definition of sampling error is taken from McBride (2016):

> *"The difference between a sample statistic used to estimate a population parameter and the actual, but unknown, value of that parameter. Here "error" does not imply that there has been a mistake; it is a technical term in statistical parlance relating to accuracy."*

In the context of the NPS-FM, the sample statistic is the NOF summary statistic (e.g., median) used to estimate the numeric attribute state. This statistic is an estimate of the 'true' attribute state because it is always based on a limited number of measurements made over a finite assessment period. There will always be uncertainty associated with estimated attribute states. Unlike environmental variability, this uncertainty can be reduced by gathering more data. A reduction in sample error leads to greater precision in the estimate of attribute state.

Increasing sampling frequency is one way to gather more data and reduce sampling error. The sampling frequency requirements associated with attributes in the NOF largely align with established or recommended council state of the environment (SOE) monitoring. The NOF requires councils to measure most water quality attributes monthly over a five-year assessment period. Other attributes such as the MCI are estimated from annual measurements only. For a few attributes, such as dissolved oxygen as 1-day and 7-day mean minima concentrations, attribute state must be estimated from near-continuous measurements collected over a period of multiple months.

The NPS-FM NOF attribute tables tend to specify a minimum number of measurements (samples) for assessing attribute state (and for some attributes, such as ammonia toxicity, no sample numbers are specified). This means, for example, that (where available) high frequency or event-based sampling could be used to better characterise the 95th percentile state. For councils that have high frequency water quality data available for some sites, it is possible to look at the differences in estimated attribute state that are based on different sampling frequencies (e.g., daily versus monthly). However, most councils only have monthly data and assessing these monthly data should be the immediate focus given the time constraints for establishing baseline attribute states.

Uncertainty associated with sampling error is commonly quantified by attaching a measure of precision, such as a confidence interval (CI) to the numeric sample statistic. However, because rivers and lakes are non-stationary systems, statistical summaries of attribute states such as medians, means and standard deviations may vary over time if long-term trends are present, or if attribute assessment periods are shorter than the periods of cyclical fluctuations (as illustrated in Figures 2-4 to 2-6). The use of CIs and other inferential statistics requires that the samples comprising the time series are independent and that the distribution is stationary (e.g., the statistical properties of the time series do not change over the assessment period). In the case of NOF attribute time-series, these requirements are likely to be violated, as indicated by long-term trends in Figures 2-4 to 2-6, and seasonal and interannual fluctuations. The use of CIs is revisited in subsection 3.2.1.

In summary, sample error will always contribute to some of the uncertainty associated with estimates of attribute state. Collecting more data can reduce sample error and lead to greater precision in attribute state estimates. If inferential statistics such as CIs are used to characterise the precision of attribute state estimates, several requirements must be met. NOF attribute time-series are likely to violate these requirements, as illustrated by the Horizons time-series plots.

2.5 Should councils implement clause 3.10(4)?

It is optional under clause 3.10(4) to account for natural variability and sampling error (i.e., sources of uncertainty) when expressing attribute states. This means that a council could decide to express attribute states as single numbers. However, scientists and resource managers have a responsibility to identify and communicate uncertainty to decision makers.

Reporting some expression of uncertainty around numeric BAS and CAS values is important in the context of subsequent NOF steps – in particular, clause 3.11(2) (setting a TAS "at or above" the baseline state), and clauses 3.18 to 3.20 (monitoring and reporting on progress to 'maintain or improve', and taking action where degradation in attribute state is identified – refer Figure 2-1). Expressing uncertainty is also important where a TAS has been set to "maintain" a BAS. For example, when attribute values fluctuate in long-period cycles (as shown in Figure 2-4), successive estimates of current state may give the appearance that attribute state is not being maintained, when in fact the CAS summary statistic (e.g., median) fluctuates over time.

In terms of the TAS, although a few panel members thought it was appropriate to express a 'range' around the threshold, the majority of the panel considered it was easier to simply treat the TAS as a single numeric threshold.

Reporting uncertainty can complicate the picture for many people, including decision makers, planners and the community. In our experience, confidence intervals or data ranges associated with numeric estimates of attribute states can also be interpreted by non-scientists as allowing for 'headroom' and for further degradation to occur.[7] However, failure to acknowledge uncertainty around numeric estimates could lead to the incorrect conclusion that there is an environmental problem (i.e., failure to meet TAS) or vice versa. Ultimately decision makers must decide whether to take a precautionary, permissive or even-handed approach to the risk of incorrect conclusions.[8] In our view, expressing uncertainty should not be seen as a means to delay actions to address degradation but rather to inform decision makers. However, there is currently no simple, rigorous statistical method for characterising uncertainty in attribute states based on time-series data.

We conclude that councils should express uncertainty in estimates of numeric attribute state provided to decision makers, but we recognise that there is no simple, rigorous statistical method to do so. In terms of BAS estimates, whether the uncertainty is expressed alongside the numeric value in a regional plan or in a background supporting technical document is a decision for council planners to make but this information should be documented and publicly available.

[7] As an example, in setting souble inorganic nutrient limits as part of their recommendations on Te Waikoropupū Springs Water Conservation Order, the Special Tribunal was uncomfortable recommending limits above current median estimates. See Microsoft Word - FINAL_Recommendation_Report_WITH_ERRATUM_20_March_2020.docx (epa.govt.nz)

[8] See McBride (2016) for a description of these three approaches. The precautionary approach assumes that a NOF 'threshold' has been breached unless the data become sufficiently convincing to indicate otherwise, the permissive approach assumes the opposite and the even-handed approach takes the data at 'face value', ignoring sampling error.

3 Implementing clause 3.10(3) and 3.10(4)

Despite the statistical problems noted in Section 2, regional councils must establish and communicate attribute states within tight timeframes, and that necessitates the provision of some preliminary advice to support the process.

We begin with commentary on the length of the assessment period(s) to adopt when calculating a BAS from existing monitoring data under clause 3.10(3). We then address how a BAS and CAS might be expressed in a way that acknowledges that these are only estimates and are influenced by environmental variability and sampling error (i.e., there is uncertainty associated with them). We follow this with examples to illustrate how councils might:

1. Estimate a BAS and characterise the associated uncertainty.
2. Set the TAS, taking the BAS into account.
3. Estimate the corresponding CAS and characterise the associated uncertainty.
4. Determine whether the CAS is on track to meet the TAS (which in the case of managing to "maintain", may also be the BAS).

We conclude this section with a brief discussion of some additional considerations when interpreting and reporting attribute states and uncertainty. These include the importance of considering multiple lines of evidence when assessing if a CAS is on track to meet the corresponding TAS. Some areas for further research are also briefly outlined.

3.1 Baseline attribute state assessment periods

The NPS-FM does not specify the time period or number of samples to use in estimating a BAS. Five years of measurements, where sampling is monthly, is often applied to water quality state assessments in freshwater state and trend reports (e.g., Whitehead et al. 2022) and provides a sufficient number of data points on which to calculate a statistically robust median (and, where applicable, 95th percentile) (e.g., McBride 2016, WHO 2003). For some attributes, the NOF also requires the CAS to be estimated from five years of measurements.

We recognise that some councils may not have five years of monitoring data for some attributes. We suggest that no less than three years of measurements are used to calculate numeric expression of a BAS from monthly measurements (based on commentary in McBride 2005 and 2016). This will enable councils to make use of slightly shorter data records or to discard earlier data impacted by a known change in the environment (e.g., recent removal/reduction of a point source discharge) or method change (e.g., arising from adoption of NEMS methods).

The "cyclic" fluctuations in Horizons Regional Council attribute time-series examples (Figures 2-4 to 2-6) suggest that, for some attributes at least, the baseline assessment period (i.e., length of time) over which a BAS is estimated may be as important as the number of measurements. Using a long assessment period might be an option to capture more of the cyclic environmental variability. However, some councils do not have long records (e.g., newly established monitoring sites or attributes). Also, estimating a BAS over a long assessment period increases the risk that the estimate will be affected by changes in land use, sampling or measurement method and other extraneous factors. In other words, there are tradeoffs between using long and short assessment periods.

3.2 Attribute state estimates and uncertainty

As a minimum, estimates of both baseline and current attribute states should be expressed as the relevant NOF summary statistics where sufficient data exist, together with their corresponding NOF attribute band and narrative description. Where councils have no or limited data to calculate or otherwise estimate baseline state, a BAS could be expressed less precisely, such as in the form of a measurement range .[9] Where this less precise BAS estimate is communicated as a NOF band or NOF band range, then it should be clearly documented that it is a coarse or interim estimate based on best available information and has not been calculated in accordance with the NOF's minimum sampling requirements. Narrative descriptions, such as those associated with the NOF bands, may also be useful for expressing baseline attribute states when data are limited.

In most cases a TAS should be expressed as a single numeric threshold. However, there will be some situations, such as where confidence in the corresponding BAS estimate is very low or what may be achievable into the future is particularly uncertain, where it may be appropriate to express an interim TAS in the form of a NOF band, rather than a specific numeric value.

Both BAS and CAS estimates should be expressed in a way that illustrates there is uncertainty associated with these estimates (associated with environmental variability and sampling error). While the TAS threshold does not require an associated estimate of uncertainty, where TAS is set to maintain BAS, it is important to recognise that future environmental variability may differ from current variability. This means that there will be uncertainty in deciding whether TAS has been achieved (owing to uncertainty in both the calculated CAS and BAS estimates). Most panel members agreed that councils could potentially accommodate some of this uncertainty when developing methods for assessing whether a TAS has been achieved (e.g., 80% of sites within an FMU might require 80% of their measurements to fall above the threshold).

3.2.1 Expression of uncertainty

We have already established that there is not a robust technical solution to quantifying uncertainty in BAS or CAS estimates arising from natural variability and sampling error, when the corresponding attribute measurements are trending upwards or downward over time, regularly fluctuating or are otherwise serially dependent. Until a better alternative can be found, we recommend that councils provide at least a narrative or qualitative description of their confidence (e.g., "low", "moderate", "high") in BAS and CAS estimates as a way of acknowledging uncertainty. If a numeric estimate of uncertainty is required, a confidence interval (CI) can be used if the time-series of attribute measurements meets the requirements for inferential statistics.

It is important to note that CIs are not descriptive statistics used to characterise datasets composed of multiple samples, such as the median, mean and variance. Rather, CIs are inferential statistics used to make predictions about the populations from which samples are taken. Specifically, a CI refers to the probability that a population parameter (e.g., a population median or 90th percentile) falls between a higher and lower value in a certain proportion of measurements (or observations), if some requirements are met. If those requirements are not met, the calculated CI is not reliable or accurate. The main requirements for CIs are that successive samples are independent and the population distribution is stationary (e.g., the statistical properties such as median and mean are constant over the assessment period). The independence and stationarity requirements are likely to be violated when using NOF attribute time-series. As noted above, these time series are often

[9] Where no data exist, this may be an expected or predicted range generated or extrapolated from modelling at other sites with similar catchment properties (e.g., geology, land use, slope) and/or expert opinion.

characterised by monotonic trends seasonal and interannual fluctuations, as illustrated in Figures 2-4 to 2-6. These temporal patterns are indicative of non-stationary populations. Non-stationarity is almost inevitable in aquatic attribute time-series, due to the nature of environmental controls such as seasonal and interannual climate cycles, and progressive land-use intensification (Hirsch et al. 2010, Snelder et al. 2022a, 2022b). For this reason, analyses of attribute time-series generally focus on identifying trends, seasonal cycles and other components of temporal variation rather than estimating the central tendency and its confidence over the entire assessment period.

If a CI is to be used to characterise uncertainty associated with a BAS or CAS, the corresponding time-series needs to meet the requirements listed above, or the data must be pre-processed to meet those requirements prior to calculating the CI. Basic pre-processing includes removing trends and seasonal variation, and ensuring that the detrended and deseasonalised data are not strongly serially correlated. Methods and software for pre-processing time-series data are widely available (e.g., Darken et al. 2002, Wu et al. 2007, Venables and Ripley 2013). Note that these pre-processing steps do not help to distinguish between human-induced versus natural trends and fluctuations.

As an alternative for communicating variability in attribute states at monitoring sites, we recommend plotting rolling estimates of attribute state over the period of record. For attribute states based on medians, a rolling window function can be used to estimate median values for a window of several years duration, shifting forward through the time series in smaller time increments. Examples of this approach are shown in Figures 2-4 to 2-6. Methods and software for computing and plotting rolling statistics are widely available (e.g., Nielsen 2019, Dama and Sinoquet 2023). This graphical approach can provide information about variation in attribute state over time, provided the time period is long enough to yield multiple windows.

We illustrate a mix of the options above in the examples that follow in subsection 3.4.

Appendix A sets out how to calculate CIs and Appendix B sets out tolerance intervals (TI) as a possible alternative statistical expression of precision. Both appendices are based on material provided by Horizons Regional Council science staff. Council staff should be explicit in their documentation that a CI provides an incomplete description of uncertainty in an attribute state because it only represents a portion of the true long-term environmental variability associated with attribute state.

3.2.2 Uncertainty assessment period

The time period over which to evaluate uncertainty in a BAS estimate is best left to the discretion of councils because we cannot recommend one time period that will be appropriate in all cases. As noted above, the NPS-FM only provides end dates for baseline assessment periods, not start dates. Some councils have long time-series for some attributes and these time-series should be assessed to identify trends or long-term cyclic fluctuations. Examining the long-term data at the outset of BAS establishment also provides an opportunity for councils to document where the estimated BAS sits within a trend trajectory and/or cycle. This provides important context for future evaluations of a CAS against the corresponding TAS and may assist with informing management intervention or re-evaluations of the TAS. It is important to note that, for non-stationary time-series, the BAS cannot be extrapolated to periods prior to or after the assessment period).

For some NOF attributes, the NPS-FM specifies the time period for assessing current attribute states; where it does not, we recommend an assessment period of five years (for the reasons outlined in subsection 3.1). However, as illustrated in the workflows in subsection 3.4, at times it will be necessary to examine the entire time-series record to provide context for where the CAS summary statistic and its associated uncertainty estimate lie within any longer-term climate-related trend and/or cycle of variation.

3.3 Reporting attribute state information to communities

In our experience with NPS-FM implementation to date, it is easier for the public to understand and digest attribute state information when it is presented in the form of categorical NOF bands and narrative descriptions rather than tables of numbers, at least in the first instance. This is particularly the case when the information spans a large number of monitoring sites and attributes. Information on uncertainty in attribute state arising from environmental variability and sampling error could be provided in the form of NOF band ranges (e.g., Band A(A-B), indicating a median of Band A with the range spanning Bands A to B), or a narrative description of confidence in the estimate of attribute state ("very likely", "highly likely", etc.).[10]

To assist with reporting progress in TAS attainment to the community, councils could explore the use of coloured 'traffic light' type flags (as illustrated in Figures 3-1 and 3-2 in the next subsection) in the first instance to indicate whether a CAS appears to be on track to meet the corresponding TAS. The CAS estimate for every NOF attribute could be presented using a square divided into green, yellow, orange and red quadrants annotated with the percentage of attributes that fall into each. This could be used to indicate problem areas that are worthy of closer attention with more detailed summary statistic information provided.

Although numeric attribute state and uncertainty estimates would not be a primary focus of summary-level attribute state reporting for communities, this information should be documented for decision makers and publicly available for those wishing to understand the state of rivers and lakes at a more detailed and technical level.

3.4 Examples

In this subsection, we set out some worked examples of the suggested approaches described above for estimating baseline and current attribute states, characterising uncertainty, and setting target attribute states. These suggestions are a starting point for further discussion and testing by regional council practitioners and may need modification to ensure that they are fit for purpose.

For simplicity, the worked examples are based on single attributes at single sites, using consistent sampling and measurement methods. In reality, a council may use data from multiple sites in an FMU to establish a BAS, TAS or CAS, and sampling and/or measurement methods may change over time. Some brief commentary on these matters is provided in subsection 3.5.

In the examples that follow, summary statistic refers to the NOF attribute statistic (e.g., median) that is calculated or otherwise estimated from available data. Where a CI is referred to, it is assumed that the requirements for using CIs have been met (refer subsection 3.2.1).

3.4.1 Expressing a baseline attribute state and associated uncertainty

Two approaches are outlined below addressing situations where sufficient data to calculate a BAS summary statistic do and do not exist. As per subsection 3.1, sufficient data here generally means a

[10] The narrative descriptions used could still reflect a statistical probability-based estimate of uncertainty (e.g., in the way confidence about the direction of a trend is currently calculated in temporal trend assessments – see Table 3-1 in subsection 3.4.3). Detailed guidance on communicating uncertainty associated with NPS-FM implementation is provided by MfE (2018).

minimum of 3-5 years of data for attributes measured monthly[11] and 4-5 years of data for attributes measured annually, if data requirements are not specified in the NPS-FM 2020.

Situation 1: Where sufficient data exist

Approach: Calculate a BAS as a numeric summary statistic and report that value along with the NOF band in which it sits (and accompanying narrative). Also include a numerical and NOF band expression of precision (noting that the measurement range could replace a precision estimate where measurements are made annually).

Where a longer-term record exists (e.g., 10+ years), plot attribute state as a rolling 3 to 5 year statistic (as per the Horizons time-series examples in Figures 2-4 to 2-6) to identify long-term trends or cyclic fluctuations.

Example 1: Total nitrogen (TN) in a polymictic lake
The lake has been sampled monthly for five years prior to the BAS establishment date adopted from the NPS-FM. The median is calculated as 480 mg/m^3 (B band) and the 90% CI is ± 30 mg/m^3, which means the CI spans both B and C bands (as 500 is the B/C boundary). Therefore, for reporting purposes: *TN BAS = 480 ± 30 mg/m^3 as a numeric; and B(B-C) as a NOF band and band range.*

Example 2: Macroinvertebrate community index (MCI)
A five-year time-series exists of MCI scores determined from annual sampling of macroinvertebrates at a river site. The median MCI score derived from the five samples is 116 with the individual numeric scores ranging from 102 to 120 (i.e., from NOF Band B to Band C). Therefore, the BAS might be reported as follows: *MCI BAS = 116 (102-120) as a numeric; and Band B(B-C) as a NOF band and band range.*

Situation 2: Where data are insufficient

This situation applies when Situation 1 does not (i.e., where *n* is <30 for attributes measured monthly (or *n* <4-5 for attributes measured annually) over the last 5 years.

Approach: Whether or not to calculate BAS as a summary statistic and/or express BAS as a NOF band and/or NOF band range(s) is a decision left to the discretion of councils. As part of making this decision, a council should consider how well the available data or other information can be translated to an indicative summary statistic, NOF band and NOF band range recognising that these are based on summary statistics (i.e., the number of available measurements is important). In some cases, a council may decide only to note the number of measurements available and their range. Where a council elects to present NOF summary statistics and/or bands, these will need to be accompanied with a note that they are indicative only (i.e., estimated from limited data).

Example 1: Chlorophyll *a* (Chl-*a*) in a default class river (as defined by the NPS-FM)
The river site has been sampled monthly for one year and on a handful of occasions in summer in two other years prior to the BAS establishment date adopted from the NPS-FM. The 92nd percentile NOF summary statistic and majority of individual sample results fall in Band B, with a range from 40 mg/m^2 (Band A) to 160 mg/L (Band C). Therefore the Chl-*a* BAS might be reported as: *Chl-a BAS* = Band B(A-C) using the NOF bands (or 40-160 mg/m^2 if the numeric range of measurements is used), where the asterisk denotes the BAS has been estimated from limited data.*

[11] Monthly sampling equates to 36 samples over three years but it is common for one or sometimes two sampling events to be missed in a year (e.g., storms or lambing that may limit access to monitoring sites). A minimum of 30 samples is consistent with the recommendations of McBride (2005) for calculation of 95th percentile metrics.

Example 2: Macroinvertebrate community index (MCI)
Three MCI scores (95, 104 and 121) are available from annual sampling of macroinvertebrates at a river site. The median of the three sample results is 104 (Band C) and the individual numeric scores span Band B and Band C. The MCI BAS estimate might therefore be reported as *MCI BAS* = 104 (if using the median), 95-121 (if using the measurement range), or Band C(B-C) (if using NOF bands), where the asterisk denotes the BAS has been estimated from limited data.*

3.4.2 Setting target attribute states

Two types of situations exist here depending on whether a decision is made to maintain or improve on the BAS.

Situation 1: Where a decision is made to set a TAS to "maintain" (i.e., TAS = BAS)

This situation can be divided into two types where the point of difference is whether or not sufficient data were available to calculate a BAS as a NOF summary statistic versus estimating a BAS only as a NOF band. In either case, where TAS is set to maintain the BAS, this means by definition "no deterioriation" from basline state and the focus can shift to evaluating trends in CAS over time (subsection 3.4.3).

Situation 1A: Where a BAS was calculated as a numeric summary statistic

Approach: Set the TAS as a numeric threshold equal to the BAS summary statistic and report that with the NOF band in which it sits (and accompanying narrative description).

Example: TN in a polymictic lake
Using the earlier example presented in subsection 3.4.1, the TN BAS was 480 ± 30 mg/m^3 or B(B-C) expressed as a NOF band and band range. Setting the TAS to maintain the BAS may be expressed as 480 (B).

Situation 1B: Where a BAS was estimated as a NOF band

Approach: Set the TAS at the same NOF band as the BAS estimate. Depending on the attribute and potential variability anticipated in measurements, the TAS might be best established as an interim target. It may also be useful for planners to consider including an explanation in the regional plan that the policy intent in this situation is to at least maintain baseline state (which may be better estimated in future) and not to allow deterioration within a NOF band due to resource use.

Example: Chl-*a* in a default class river
Using the earlier example presented in subsection 3.4.1, the chl-*a* BAS was estimated to be in NOF Band B. Therefore setting a TAS to maintain the BAS might be described as B (40-160 mg/m^2) (if the council elects to report the numeric measurement range for the BAS estimate with the indicative NOF band).

Situation 2: Where a decision is made to set a TAS to "improve" relative to the BAS

There are several options to set a TAS as an improvement in attribute state relative to the corresponding BAS. For example, an improvement could be sought in the median attribute state and/or in the form of fewer measurements at the poor end of the range of measurements. The concept of seeking an improvement in a proportion of measurements is reflected in Appendix 3 of the NPS-FM (national target for primary contact recreation).

Situation 2A: Where a BAS was calculated as a numeric summary statistic

Approach: Set the TAS as a numeric threshold lower (or higher in the case of an attribute such as dissolved oxygen concentration) than that of the BAS NOF summary statistic and report this alongside the NOF band in which the TAS sits (and accompanying narrative description).

Example: TN in a polymictic lake
Once again using example 1 in subsection 3.4.1, TN has a BAS (as a median) of 480 ± 30 mg/m^3 or B(B-C) expressed as a NOF band and band range. If the aim is to improve:

- within the B band (which spans from 300-500 mg/m^3) from the current near bottom of Band B to nearer to the top of Band B (say 350 mg/m^3) then the TAS might be expressed as a median of 350 mg/m^3 (B).
- from the current B band to the A band, then the TAS might be expressed as a median of <300 mg/m^3 (A).

Situation 2B: Where a BAS was estimated as a NOF band

Approach: Set the TAS so that it is at a higher NOF band or improved measurement or NOF band range than the BAS estimate. Express this TAS as a band(s) together with its upper and lower numeric boundaries (and accompanying narrative description).

Example: Revisiting example 2 of Situation 2 from subsection 3.4.1, Chl-*a* in a default class river has a measurement range (as a 92nd percentile) of 40-160 mg/m^2, and spans Band B(A-C) (0-200 mg/m^2) based on the NOF bands and their numerical range. If the aim is to improve Chl-a to:

- reduce the upper end of the measurement range by 50%, then the TAS might be expressed as a 92nd percentile of B and <80 mg/m^3, or
- sit consistently within the B band (which spans from 51-120 mg/m^3), then the TAS might be expressed as a 92nd percentile of B (51-120 mg/m^3), or
- at least the B band, then the TAS might be expressed as a 92nd percentile of A-B (<120 mg/m^3).

3.4.3 Estimating a CAS (and an expression of associated uncertainty) and assessing progress towards the corresponding TAS

At the point in the NOF process where a CAS is being calculated, the requirements of clause 3.18 (monitoring progress towards achieving TAS and environmental outcomes) and clause 3.19 (assessing trends in attribute states and investigating any deteriorating trend that is "more likely than not") of the NPS-FM come into play. The workflows below therefore combine both an assessment of CAS and temporal trends.[12] For assessment and interpretation of temporal trends, we recommend councils follow the guidance of Snelder et al. (2021).

In the example workflows summarised in Figures 3-1 and 3-2, a deteriorating trend is defined as one in which a deterioration is at least "likely", equating to a probability between 0.67 and 1 (Table 3-1). Traffic-light style colour coding can be used to support visual interpretation of the trend results. Again, two types of situation exist here based on whether a TAS was set with the aim of maintaining or improving on its corresponding BAS.

[12] Clause 3.20 (responding to degradation) is also important here as it is inextricably linked to clause 3.19. When applying clause 3.20, councils should use information on trend direction and magnitude – and the associated confidence in these estimates – to decide whether to act and, if so, how strongly to act (other factors, such as the potential consequences of not acting are also relevant).

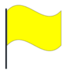
OUTCOME 1

The CAS summary statistic (i.e., numeric NOF summary statistic) is <u>better than or equal to</u> the TAS numeric threshold and no deteriorating temporal trend is identified

Council is currently meeting the TAS; continue with routine monitoring and reporting.

OUTCOME 2

The CAS summary statistic is <u>better than or equal</u> to the TAS numeric threshold but a deteriorating temporal trend is identified*



A potential issue may exist; check confidence in trend direction and the magnitude of trend. Where long-term time-series data are available, check where the CAS estimate sits within any longer-term climate-related trend and/or cycle of variation and document and communicate this. Continue to maintain a watching brief via routine monitoring and reporting.

**A different outcome that might also attract a yellow flag is when the CAS summary statistic is lower (poorer) than the TAS numeric threshold but falls within the same NOF band (regardless of the temporal trend assessment results).*

OUTCOME 3

The CAS summary statistic sits within a lower (poorer) NOF band than the TAS numeric threshold but an improving temporal trend is identified

The CAS is not meeting the TAS but the trend direction is improving; check confidence in both trend direction and magnitude to determine the extent of investigations necessary. Evaluate other lines of evidence such as changes in catchment land use/management, climate cycle/hydrological events, degradation in related attributes, etc.). Additional monitoring may also be necessary.

OUTCOME 4

The CAS summary statistic sits in a lower (poorer) NOF band than the TAS numeric threshold and there is no evidence of an improving trend of sufficient magnitude to meet the TAS by the target date

The CAS is not meeting the TAS and the associated temporal trend is identified as (a) deteriorating, (b) as likely to be deteriorating as improving, or (c) improving at an insufficient magnitude. Evaluate both the confidence in any deteriorating trend and the magnitude of the deterioration to inform council decisions around regulatory and non-regulatory interventions (incl. additional monitoring). Include a detailed investigation that considers multiple lines of evidence (e.g., changes in catchment land use/management, climate cycle/hydrological events, degradation in related attributes or sites) and comment on the potential consequences of the observed degradation for the freshwater environment in question.

Figure 3-1: Example outcomes for assessments of a CAS against its corresponding TAS where the TAS was set with the aim of maintaining the BAS. For clarity, the TAS numeric threshold refers to the summary statistic numeric calculated for its corresponding BAS in subsection 3.4.1; either a specified numeric target if that has been formally identified, or the bottom of the NOF band that has been identified as the target band if a numeric has not been identified.

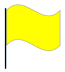
OUTCOME 1

The CAS summary statistic statistic (i.e., numeric NOF summary statistic) is <u>better than</u> the TAS numeric threshold and no deteriorating temporal trend is identified

Council is currently meeting the TAS; continue with routine monitoring and reporting.

OUTCOME 2

The CAS summary statistic is <u>better than</u> the TAS numeric threshold but a deteriorating temporal trend is identified*



A potential issue may exist; check confidence in trend direction and the magnitude of trend. Where long-term time-series data are available, check where the CAS sits within any longer-term climate-related trend and/or cycle of variation and document and communicate this. Continue to maintain a watching brief via routine monitoring and reporting.

**A different outcome that might also attract a yellow flag is when the CAS summary statistic is lower (poorer) than the TAS numeric threshold but falls within the same NOF band (regardless of the temporal trend assessment results).*

OUTCOME 3

The CAS summary statistic sits within a lower (poorer) NOF band than the TAS numeric threshold but an improving temporal trend is identified

The CAS is not meeting TAS but the trend direction is improving; check confidence in both trend direction and magnitude to determine extent of investigations necessary. Evaluate other lines of evidence such as changes in catchment land use/management, climate cycle/hydrological events, degradation in related attributes, etc.). Additional monitoring may also be necessary.

OUTCOME 4

The CAS summary statistic sits in a lower (poorer) NOF band than the TAS numeric threshold and there is no evidence of an improving trend of sufficient magnitude to meet the TAS by the target date

The CAS is not meeting the TAS and the associated temporal trend is identified as (a) deteriorating, (b) as likely to be deteriorating as improving, or (c) improving at an insufficient magnitude. Evaluate both the confidence in any deteriorating trend and the magnitude of the deterioration to inform council decisions around regulatory and non-regulatory interventions (incl. additional monitoring). Include a detailed investigation that considers multiple lines of evidence (e.g., changes in catchment land use/management, climate cycle/hydrological events, degradation in related attributes or sites) and commentary on the potential consequences of the observed degradation for the freshwater environment in question.

Figure 3-2: Example outcomes for assessments of a CAS against its corresponding TAS where the TAS was set with the aim of improving the BAS. For clarity, the TAS numeric threshold refers to the summary statistic numeric calculated for its corresponding BAS in subsection 3.4.1; either a specified numeric target if that has been formally identified, or the bottom of the band that has been identified as the target band if a numeric has not been identified.

Table 3-1: Confidence categories for conveying the likelihood (probability) of a deteriorating trend in a freshwater attribute. The confidence categories presented here are a simplified version of those used by the Intergovernmental Panel on Climate Change (IPCC; Stocker et al. 2014). See Snelder et al. (2021) for more commentary.

Probability (%)	Level of confidence in trend and its direction
0–10	Highly unlikely
10–33	Unlikely
33–67	As likely as not to have increased or decreased
67–90	Likely
90–100	Highly likely

For simplicity, the workflows assume that there are sufficient measurements (or other information such as modelled estimates) upon which to calculate the CAS as a numeric NOF summary statistic (e.g., median) with an associated estimate of precision (e.g., 90th percentile CI, calculated over the CAS assessment period). Both the CAS summary statistic and estimate of precision should be included in council reporting, with the latter included only for the purpose of indicating the amount of variability (uncertainty) in the CAS estimate.

In the workflows, the results of a temporal trend assessment should be used to identify if CAS is on track to meet TAS. This is discussed further in subsection 3.5.

Situation 1: Where a TAS was set with the aim of maintaining the BAS

Approach: Calculate the CAS summary statistic according to the NOF attribute table requirements (if not specified, apply the "sufficient data" requirements as per subsection 3.1) and carry out a temporal trend assessment. Follow the advice in the relevant assessment outcome box of Figure 3-1. Note that Figure 3-1 is intended to provide generic guidance only; other outcomes may be possible.

Situation 2: Where a TAS was set with the aim of improving on the BAS

Approach: Calculate as per Situation 1 but follow the advice in the relevant assessment outcome box of Figure 3-2.

3.5 Discussion

Implementation of clause 3.10(3)-(4) of the NPS-FM is not straightforward and this has flow-on effects for attribute state assessments and reporting under clause 3.30(2) that may not have been anticipated when the NPS-FM 2020 was prepared.

Including a statistical estimate of precision with estimates of BAS and CAS-provides a way to indicate a component of uncertainty in these attribute states. However, this estimate is not robust when attribute measurements come from non-stationary populations, or fluctuate seasonally, or if there is auto-correlation in the measurement time-series. Further, this leaves unquantified components of uncertainty (such as that arising from environmental variability beyond a given assessment period) that will confound comparisons between different attribute states. For example, recent work by Snelder and Kerr (2022) for Auckland Council on the influence of flow on water quality in rivers has demonstrated that assessments of attribute state (based on five years of monthly measurements) were influenced by differences in the flow regime between assessment periods (which were in turn likely due to climate variability that may have been natural or associated with anthropogenic climate change, or both). This work exemplifies the difficulty in determining whether changes in attribute state between assessment periods are due to manageable human activities or unmanageable natural processes.

Given that it is not currently possible to make rigorous statistical comparisons of attribute state between different assessment periods, we recommend that temporal trend assessments are used as the primary means to indicate whether a CAS is on track to meet its corresponding TAS. We recognise that temporal trend assessments also involve assumptions and, therefore, there is uncertainty associated with trend direction and magnitude estimates. In addition, statistical trends cannot be extrapolated beyond the end date of an attribute state assessment period, which precludes inferences about ‘over-shooting’ or ‘under-shooting’ the TAS in the future. However, standard trend assessment procedures are already well established (see Snelder et al. 2021) and there are techniques available to potentially account for variability associated with seasonal cycles and flow fluctuations (e.g., Weighted Regressions on Time, Discharge and Season (WRTDS)).[13]

The workflows in Figures 3-1 and 3-2 illustrate how trend assessments could be used by councils to decide whether further assessments or actions are needed. If the trend directions indicate anything other than green flags, then more information should be evaluated, including:

- the entire time-series record to see whether the numeric CAS estimate and its associated expression of uncertainty lie within the longer-term climate-related trend and/or cycle of variation,
- the statistical confidence associated with the temporal trend estimate, and
- temporal trend magnitude (and the associated confidence in this estimate).

Contextual details that should be provided to decision makers tasked with determining if management actions are needed to meet a TAS should include:

- whether there has been a change in catchment land use and/or management (which will require monitoring of the implementation of relevant regional plan provisions and non-regulatory initiatives),
- whether a change in attribute state or trend is also evident in a related attribute (e.g., if a deteriorating state or trend is observed in nitrate toxicity, is there an associated deterioration in the macroinvertebrate attribute state summary statistics that may suggest an adverse impact on ecosystem health?) or at unimpacted/reference sites within the same area/catchment (if so, then management intervention may not be required),
- whether long-term climate cycles or specific hydrological (or other) events may have influenced attribute state over an assessment period (e.g., present CAS as a time-series graph along the lines of Figure 2-4 so that where CAS sits within the range of longer-term cyclic variability can be seen), and
- whether any changes in sampling and/or measurement methods might have impacted the attribute assessment (see subsection 3.5.3).

[13] Note however that the WRTDS method does not necessarily remove the influence of climate on attribute time-series records.

3.5.1 Attribute states from multiple sites (spatial variability)

Although spatial variability was not considered in this report, we recognise that a council may use data from multiple sites in a FMU to establish attribute states. This approach ensures that estimates of attribute state capture some spatial variability that will likely exist across a FMU. In terms of expressing an attribute state across multiple sites, we recommend that the numeric NOF summary statistic or NOF band and some associated estimate of uncertainty are also documented for each individual site. The council will then need to derive a way to assign an overall classification of attribute state from those individual numerics. One way this could be done is to report the median and ranges (where available) and associated NOF band and numeric band range across the sites.

<u>Example:</u> Chlorophyll *a* (Chl-*a*) in a default class river
Data are available for five river sites within an FMU and a council wishes to establish the BAS. The individual 92% percentile values from monthly sampling for the previous three years are:

- River/site 1: 40 mg/m^2 (Band A)
- River/site 2: 100 mg/m^2 (Band B)
- River/site 3: 85 mg/m^2 (Band B)
- River/site 4: 180 mg/m^2 (Band C)
- River/site 5: 110 mg/m^2 (Band B)

The median value is 100 mg/m^2, so the BAS estimate might be expressed as: *Chl-a BAS = Band B(A-C) (100; 40-180 mg/m^2).*

If a TAS was set to maintain baseline state (Situation 1 described in subsection 3.4.2), then the procedure used in the example shown above could also be used to set a TAS for an FMU. Several panel members suggested that uncertainty about the achievement of a TAS could be addressed by establishing at the outset the percentage of monitoring sites – or potentially monitoring years – that need to meet the TAS threshold for the FMU-wide TAS to be deemed to have been met (in the example above, say 80% (4 out of 5) sites must have a 92% percentile biomass of <100 mg/m^2 or possibly that the 92% percentile just falls in Band B). Another potential expression of TAS that might warrant further investigation is as a percentage improvement that is required to be achieved at a site or across multiple sites (e.g., the median concentration at site X, or the median concentration at 80% (8 out of 10) sites, must decrease by at least 20%). This approach might assist with the application of on-farm mitigations and catchment water quality models, for which relative changes may be more appropriate than absolute values.

3.5.2 Changes in sampling and/or measurement methods

Sampling and/or measurement methods (including method detection limits) inevitably change over time, such as through adoption of NEMS protocols or improved tools/technologies. The effect of method changes on attribute measurements will need to be evaluated (e.g., through paired measurement campaigns or literature reviews) and the historic measurements adjusted where possible. If this is not possible, changes in methods should be documented to flag that the changes may influence an attribute time-series.

We recommend documenting measurement methods at the time of estimating baseline attribute state and noting pending method changes.[14] Similarly, each time a CAS is reassessed and a trend assessment performed, it will be necessary to consider and document any changes in sampling and/or measurement methods.

3.5.3 River flows

Assessments of attribute state at river monitoring sites are influenced by both the instantaneous flow rate at the time of water sampling and the longer-term flow regime prior to sampling (Snelder and Kerr 2022). Sampling, therefore, needs to be unbiased with respect to instantaneous flow to ensure that attribute measurements represent the true attribute state. Continuous flow monitoring or modelling is recommended so that the flow regime variation can be characterised and used as contextual information alongside reporting of attribute state and trends.

3.5.4 Climate

As discussed earlier in this report, climate variability can affect estimates of attribute state and trends and may cause or contribute to failures to achieve a TAS. It would, therefore, be useful to investigate methods for removing the cyclical fluctuations in attribute time-series associated with climate variability, and for evaluating the residual variation about the median attribute state.

Most of the work to date on quantifying the effects of cyclical climate processes on trends in freshwater attributes has used SOI as the sole climate indicator variable (Scarsbrook et al. 2003, Snelder et al. 2022a, b). It would, therefore, be useful to identify other proxy measures of climate that may be able to explain more of the variation observed in attribute states.

[14] For example, some councils were up until recently calculating total nitrogen (TN) from the sum of Total Kjeldahl Nitrogen (TKN) and dissolved inorganic nitrogen (DIN) but have now adopted the NEMS recommended direct persulphate TN method. Extensive paired testing has established that the TKN test, owing to a stronger acid digestion, returns higher TN concentrations from sediment-laden water samples than the direct persulphate TN method (e.g., Davies-Colley and McBride 2016) – this difference has been sufficient to result in (artificial) shifts in TLI scores in some lakes.

4 Conclusions and recommendations

This report has highlighted that a robust technical solution to implementing clause 3.10(3)-(4) of the NPS-FM does not currently exist. Clause 3(10)(4) usefully recognises that both natural variability and sampling error are sources of uncertainty in the estimation of baseline and current attribute states. However, in practice, it is very difficult to distinguish sampling error from natural variability and we do not have the ability to robustly quantify these sources of uncertainty when making assessments of baseline and current attribute states. One particularly challenging issue is how to account for long-term cyclical fluctuations in river and lake attribute states that are generally not captured in estimates of uncertainty based on relatively short assessment periods. Another challenging issue is how to account for the influence of seasonality on uncertainty in attribute states during assessment periods.

Councils need to estimate BAS and CAS and set TAS within tight timeframes, and we have provided some preliminary commentary and suggestions to support councils with these tasks. This includes consideration of the assessment period for establishing a BAS and possible approaches to characterise uncertainty in estimates of baseline and current attribute states. Councils could report attribute states without estimates of precision or variability, but we do not recommend this approach because it conveys no information about the uncertainty associated with attribute state estimation. Instead, we suggest that BAS and CAS estimates are reported with some estimation or description of uncertainty. The use of CIs to meet this need was discussed by the expert panel. The use of CIs requires that the samples comprising an attribute time-series are independent and come from a stationary distribution. In the case of NOF attribute time-series, these requirements are likely to be violated, as indicated by long-term trends and seasonal and interannual fluctuations. Under these conditions, time series data must be pre-processed to remove trends and fluctuations, and checked for serial autocorrelation before calculating CIs.

As an alternative, councils can assess temporal variability in the attribute state over the assessment period using a rolling time window (e.g., a rolling median). At the very least, councils should provide a narrative description of their confidence in the estimate of attribute state based on expert judgement.

Given the lack of readily available methods to make statistically robust comparisons of current and target attribute states, we recommend that temporal trend assessments, for which standard procedures are already well established, are used as the primary means to indicate if a CAS is on track to meet its corresponding TAS.

Decision makers will need a range of information to contextualise and interpret the results of attribute state and trend assessments. This includes evaluations of changes in catchment land use/management, changes in related attributes or reference sites, and the effects of climate cycle/hydrological events during the assessment period. This contextual information will be particularly important for decision makers tasked with determining whether management actions are needed to meet a TAS and the broader environmental outcomes established under the NPS-FM.

Methods and accompanying detailed national guidance need to be developed to support robust comparisons of attribute state between assessment periods, including accounting for natural variability and sampling error. Specific work is needed to better understand, quantify and account for the influence of seasonality, cyclical climate processes, and other drivers of the natural variability associated with attribute state (and trends) through time. For example, we recommend:

- statistical methods to remove the effect of long-term trends and seasonal and interannual fluctuations in attribute time-series data to evaluate the residual variation about the estimated attribute state,
- investigating new proxy measures of climate variability that may be correlated with variation in attribute state, and
- investigating methods to characterise variation in time-series data from non-stationary distributions.

This report focussed on temporal variability at the scale of a single site. Spatial variability forms a significant component of natural variability and national guidance is also needed on how to account for this component in the context of clause 3.10(4). Overall, given the complexities involved with accounting for natural variability and sampling error when establishing and comparing between different attribute states, it may be necessary to revisit some details of attribute state assessments and reporting specified in the NPS-FM 2020.

5 Acknowledgements

This report was prepared with an MBIE Envirolink Advice Grant. We thank Dr Ton Snelder (LWP Ltd), Dr Coral Grant (Auckland Council), Shirley Hayward (Environment Canterbury), James Dare (Bay of Plenty Regional Council) and Maree Patterson (Horizons Regional Council) for their input at the workshop. Ton also contributed to valuable discussions on aspects of this report. Dr Jenni Gadd reviewed an early version of this report and Dr Scott Larned reviewed and improved subsequent drafts.

6 Glossary

Accuracy	Closeness of agreement between measurements of an attribute (using NPS-FM language) and the attribute's true (unknown) value. Accuracy includes a combination of both precision and bias.
Assessment period	The time period used to calculate an attribute state metric (e.g., 5 years).
Bias	Consistent or systematic over- or under-reporting in the measurement of an attribute (e.g., due to a field or laboratory sensor that consistently over-reports a measurement).
Confidence interval	A statistical interval within which, with some designated confidence level (e.g., 90 or 95%), a measurement will lie most of the time, under repetitive sampling.
Coefficient of variation	A measure of variability, expressed as a percentage, derived by dividing the standard deviation of a dataset by the mean value.
Measurement	An individual sample (in NPS-FM language) or observation (in statistical language).
Non stationary	A time-series whose statistical properties such as mean and standard deviation do not remain consistent over time.
Numeric attribute state	The face value of the relevant summary statistic for a NOF attribute (e.g., a median total nitrogen concentration of 2.1 mg/L), calculated from a set of measurements over a specific assessment period.
Precision	How close repeated independent measurements are to each other.
Residual	A statistical term meaning the difference between an observed value and the corresponding value of the same attribute when predicted by a function.
Sampling error	The difference between a sample statistic used to estimate a population parameter and the actual but unknown value of the parameter (see subsection 2.4).
Sampling frequency	The frequency at which a population is sampled. In the NPS-FM, some attributes must be sampled (measured) at a specific time interval (e.g., monthly or annually).
Tolerance interval	A statistical interval within which, with some designated confidence level (e.g., 90 or 95%), a specified proportion of a sampled population falls.
Uncertainty of measurement (UoM)	An estimate of the variability inherent in a measurement based on instrument / equipment calibrations, purity of chemicals used for making calibration standards, and human factors.

7 References

Darken P.F., Zipper C.E., Holtzman G.I., Smith E.P. (2002) Serial correlation in water quality variables: estimation and implications for trend analysis. *Water Resources Research,* 38: 22-1.

Dama F., Sinoquet C. (2021) Time series analysis and modeling to forecast: a survey. *arXiv*:2104.00164.

Davies-Colley R., McBride G. (2016) Accounting for changes in method in long-term nutrient data: Recommendations based on analysis of paired SoE data from Wellington rivers. NIWA Client Report No. HAM2016-070 prepared for the Ministry for the Environment, Wellington.

Fullard L., Patterson M., Patterson M., Lennard A. (2022) The generation and interpretation of confidence intervals for state calculations. Unpublished draft memorandum, Horizons Regional Council, 20 September 2022.

Goudey R. (2007). Do statistical inferences allowing three alternative decisions give better feedback for environmentally precautionary decision-making? *Journal of Environmental Management*, 85(2): 338-344.

Helsel D.R., Hirsch R.M. (1992) Statistical methods in water resources (Vol. 49). Elsevier.

Hirsh R.M., Moyer D.L., Archfield S.A. (2010) Weighted regressions on time, discharge, and season (WRTDS), with an application to Chesapeake Bay river inputs. *Journal of the American Water Resources Association,* 246: 857-80.

Hyndman R.J., Fan Y. (1996) Sample quantiles in statistical packages. *The American Statistician*, 50(4): 361-365.

McBride G.B. (2005) Using statistical methods for water quality management: issues, problems and solutions. John Wiley & Sons, New York.

McBride G. (2016) NOF statistical considerations for design and assessment. NIWA Client report HAM16022 prepared for Ministry for the Environment. 47 p.

Ministry for the Environment (2003) Microbiological water quality guidelines for marine and freshwater recreational areas. Ministry for the Environment, Wellington.

Ministry for the Environment (2018) A guide to communicating and managing uncertainty when implementing the National Policy Statement for Freshwater Management 2014. Ministry for the Environment, Wellington.

Ministry for the Environment (2022) Guidance on the National Objectives Framework of the National Policy Statement for Freshwater Management. Ministry for the Environment, Wellington.

Nielsen A. (2019) Practical time series analysis: prediction with statistics and machine learning. O'Reilly Media, Sebastopol, CA, USA.

New Zealand Government (2020) National Policy Statement for Freshwater Management. New Zealand Government, December 2022.

Royal Society of Chemistry (2003) Terminology – the key to understanding analytical science. Part 1: Accuracy, precision and uncertainty. *Analytical Methods Committee Technical Brief No. 13*. 2 p.

Scarsbrook M.R., McBride C.G., McBride G.B., Bryers G.G. (2003) Effects of climate variability on rivers: consequences for long term water quality analysis. *Journal of the American Water Resources Association*, 39(6): 1435-1447.

Snelder T., Fraser C., Larned S., Whitehead A. (2021) Guidance for the analysis of temporal trends in environmental data. NIWA Client Report 2021017WN prepared for Horizons Regional Council and MBIE Envirolink.

Snelder T.H., Larned S.T., Fraser C., De Malmanche S. (2022a) Effect of climate variability on water quality trends in New Zealand rivers. *Marine and Freshwater Research*, 73: 20-34.

Snelder T.H., Larned S.T., Fraser C., Monaghan R., Whitehead A.L., De Malmanche S. (2022b) Attribution of river water-quality trends to agricultural land use and climate variability in New Zealand. *Marine and Freshwater Research*, 73: 1-19.

Snelder T.H., Larned S.T, McDowell R.W. (2017) Anthropogenic increases of catchment nitrogen and phosphorus loads in New Zealand. *New Zealand Journal of Marine and Freshwater Research*, DOI: 10.1080/00288330.2017.1393758

Stocker T., Qin D.Q., Plattner G.-K. (eds). (2014) Climate change 2013: the physical science basis: Working Group I Contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press.

Snelder T., Kerr T. (2022) Relationships between flow and river water quality monitoring data and recommendations for assessing NPS-FM attribute states and trends. Report prepared for Auckland Council by LWP Ltd.

Venables W.N., Ripley B.D. (2013) Modern applied statistics with S-PLUS. Springer Science & Business Media.

Whitehead A.L., Fraser C., Snelder T., Walter K., Woodward S., Zammit C. (2022) Water quality state and trends in New Zealand rivers: analyses of national data ending in 2020. NIWA Client Report No. 2021296CH prepared for Ministry for the Environment. 73 p.

WHO (2003) Guidelines for safe recreational water environments. Volume 1, Coastal and fresh waters. World Health Organization. 219 p.

Wilson E.B. (1927) Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158): 209-212.

Wu Z., Huang N.E., Long S.R., Peng C.K. (2007) On the trend, detrending, and variability of nonlinear and nonstationary time series. *Proceedings of the National Academy of* Sciences, 104: 14889-14894.

Appendix A Confidence interval calculations

The Wilson score method (Wilson 1927) for calculating the confidence interval of a percentile is described below, reproduced from Fullard et al. (2022). This method treats each water quality sample as a Bernoulli trial, that is, each sample is less than or equal to the desired population percentile (such as the median) or not. Assessing the data in such a way means that the results of our Bernoulli trial are binomially distributed, regardless of the underlying distribution of the data itself. Therefore, the generation of uncertainty intervals for percentiles is achieved using well-studied methods for binomial uncertainty intervals (Goudey 2007, Helsel and Hirsch 1992).

In essence:

- Given a desired percentile estimate, p (where $0 \leq p \leq 1$), and confidence level, $(1-\alpha)$, use the Wilson score method (Equation 1.1) to calculate the interval (p^{-}, p^{+}).
- For the ends of the interval, p^{-}, and p^{+}, and the percentile point estimate p, convert the percentile estimates into sample unit values using the Hazen percentile method.

Let:

n = the number of samples

p = the desired percentile point estimate

$z_{\alpha/2}$ = the critical z-value from the standard normal distribution (i.e., for a 95% confidence interval, $z_{0.025} = 1.96$).

$(1-\alpha)$ = the confidence level for a two-sided interval.

Given the above definitions, the Wilson score confidence interval for the percentile can be written as:

$$(p^{-}, p^{+}) = p' \mp z_{\alpha/2} s', \tag{1.1}$$

where

$$p' = \frac{\left(p + \frac{z_{\alpha/2}^2}{2n}\right)}{D'}, \tag{1.2}$$

$$s' = \frac{\sqrt{\frac{p(1-p)}{n} + \frac{z_{\alpha/2}^2}{4n^2}}}{D'}, \tag{1.3}$$

and

$$D' = 1 + \frac{z_{\alpha/2}^2}{n}. \tag{1.4}$$

Equation 1.1 allows us to estimate the confidence interval for the desired true population *percentile*. We can interpret this interval as follows: for a desired true population percentile, p, we can be $100 * (1-\alpha)\%$ confident that the true **population** percentile is between the **sample** percentiles (p^{-}, p^{+}). There is one further step required to turn this interval estimate into a confidence interval for the *input data* (i.e. to change the confidence interval for the percentile into the confidence interval for the data in the same units as the raw data). Given the lower confidence interval estimate for the percentile, p^{-}, a lower confidence limit for the input data is found by taking the p^{-}th Hazen percentile of the input data. Given the upper confidence interval estimate for the percentile, p^+, an upper confidence limit for the input data is found by taking the p^+th Hazen percentile of the input data.

While there are multiple methods for calculation of percentiles, we choose here to use the Hazen percentile method as this is popular for water quality percentile assessment (MfE 2003). Calculation details for the Hazen percentile method are found in Hyndman and Fan (1996) and involve the generation of a piecewise linear function between known order statistics (i.e., percentiles approximated from the raw data). The desired percentile is then obtained by linear interpolation using this function. In the R package "*quantile*" the Hazen percentile can be calculated by setting the "*type"* parameter to 5.

Example R-code:
Example R code for calculation of the Wilson score interval for a desired percentile with a desired confidence, and also for calculation of percentiles using the Hazen method, are provided in Appendix C.

Example 1: We take a sample of a standard normal distributed random variable with $n = 50$. If we calculate the median with 90% confidence ($\alpha = 0.1$) we find:

```
The confidence limit for the percentile =  0.5 :[    0.3867159 0.6132841
```

The above R-script output tells us the 90% confidence interval for the median (percentile = 0.5) is (0.387, 0.613). This result means that we are 90% confident that the true (population) median of the data is between the 38.7th and 61.3rd percentiles of the sample data. Therefore, to calculate the confidence interval for the median, we must take the 38.7th and 61.3rd Hazen percentiles of the sample data:

```
The confidence limits of the data: -0.07976 0.33654
```

The R-script output tells us that the 90% confidence limit for the true median of our normally distributed data is (-0.07976, 0.33654) which includes the expected mean value of 0.

The point estimate for the median is found by calculating the 50th Hazen percentile:

```
Hazen point estimate for the percentile:  0.17175
```

This is within the confidence limits defined.

Example 2: We take a sample of a uniformly distributed random variable with $n = 50$. If we calculate the 95th percentile with 90% confidence ($\alpha = 0.1$) we find:

```
The confidence limit for the percentile =  0.95 :    0.8723846 0.9814155
The confidence limits of the data: 0.91346 0.97105
 Hazen point estimate for the percentile:  0.9559
```

This gives a confidence interval which contains the expected value (0.95).

Appendix B Tolerance interval calculations

A tolerance interval is an uncertainty interval, which differs from a confidence interval in that a confidence interval gives a range around a certain statistical measure (i.e., a range of plausible values for the mean, median, other percentile), while a tolerance interval gives a range in which a specified proportion of the sample population exists. In other words, a confidence interval provides a ranged estimate around a statistical parameter, while a tolerance interval gives a range of values that exceed a given numeric value or attribute state (e.g., 90% of all possible values of the data). Like a confidence interval, a tolerance interval has an associated confidence level, α, which is a user defined choice.

Horizons has been investigating the potential use of **Wilson-Hazen tolerance limits** and consider they may be preferable to confidence intervals (and Null hypothesis significance testing or NHST). The following text is reproduced from Fullard et al. (2022).

Wilson-Hazen Tolerance Limits

A **tolerance limit** is an estimate which, with some associated confidence, tells us that a certain percentage of the data is below a value (an Upper Tolerance Limit (UTL)) or above a certain value (a Lower Tolerance Limit (LTL)). The generation of tolerance limits is achieved through the generation of a confidence interval for a specified percentage, with an associated confidence level.

For example, if we wanted to calculate a 90% upper tolerance limit ($p = 0.9$) with confidence level 95% ($\alpha = 0.05$) we can simply calculate a 95% confidence interval for the 90th percentile using methods similar to those in Appendix B and take the upper confidence limit value (p^+) as our UTL. In this way, 90% of values in our dataset are expected to be less than the UTL, p^+, with 95% confidence.

The only difference between the generation of an UTL or LTL is that a one-sided confidence interval is generated. The two-sided confidence intervals presented in Section 3.1 were necessary to bound the percentile value, but here we are interested in everything below the upper bound of the upper confidence interval estimate, hence only a one-sided limit is necessary. Using a one-sided interval has the advantage of giving greater statistical power for the same sample size compared to the two-sided limit, since all of the tail of the distribution is above the desired percentile, and not split in half at the upper and lower tails.

For our purposes, we define an (upper or lower) Wilson-Hazen Tolerance Limit as being a one-sided tolerance limit generated using the methodology of Appendix B. Given this definition, we can extend this theory to generate a probability of our current attribute state being better than a designated target attribute state (referred to as a threshold state by Horizons). Depending on whether the percentile point estimate (i.e., current attribute state) is below or above the threshold value (i.e., TAS), we calculate a series of either LTL, or UTL, with various levels of confidence. For example, Figure B-1 provides an example of where the percentile point estimate is greater than the threshold value (left) and one where the point estimate is less than the threshold value (right).

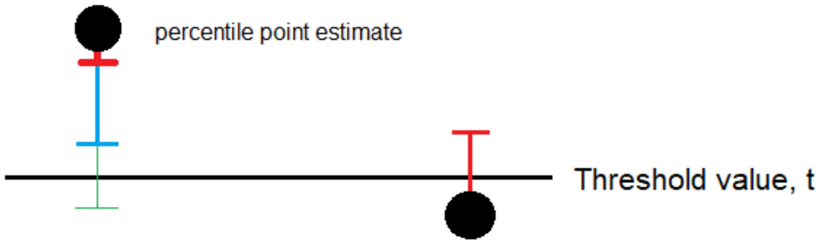


Figure B-1: Example of a percentile point estimate which is above a threshold value (left) where only the lower tolerance limit (a 95th percentile) intersects the threshold value, or a percentile point estimate which is below the threshold value (right) where the upper tolerance limit intersects the threshold value.

For the first example (left) where the percentile point estimate is greater than the threshold value we calculate a series of Lower Threshold Limits (LTL) with increasing confidence level. For example, the LTL in red may have a confidence level of 75% ($\alpha = 0.25$), the LTL in blue may have a 90% confidence level ($\alpha = 0.1$), and the one in green may have a 95% confidence level ($\alpha = 0.05$). Since the threshold is crossed between the 90 and 95% confidence levels we can interpolate the values of the LTL values to find an estimate of the confidence level *at the threshold value*. In this example, we may find a LTL with a 92% confidence level exactly meets the threshold value. <u>Therefore, we say that there is a 92% chance that our percentile is above the threshold value (an 8% chance that we are below the threshold (i.e., TAS) value).</u> The concept is identical in the second example (right) where the percentile point estimate is less than the threshold value, but in this case we would calculate a series of Upper Tolerance Limits.

To summarise, the process of generating the probability of a percentile being above or below a threshold value is as follows:

1. Given a desired percentile estimate, p (where $0 \le p \le 1$), use the Hazen method to calculate the percentile point estimate.
2. a) If the percentile point estimate is greater than the threshold value, generate a series of Lower Threshold Limit values using the one-sided Wilson-Hazen method at increasing confidence level.
   b) If the percentile point estimate is less than the threshold value, generate a series of Upper Threshold Limit values using the one-sided Wilson-Hazen method at increasing confidence level.
3. Interpolate the threshold limit values on the confidence level values to find the confidence level exactly at the threshold value.
4. Interpret this confidence value to find the probability that the percentile value is above or below the threshold value.

Example 1: We take a sample of a uniformly distributed random variable with $n = 50$ between the values of 0 and 1. If we calculate the probability that the median is less than a TAT threshold value of 0.45:

```
Threshold value =  0.45
Probability that percentile < threshold:  0.3708485
```

At p=0.37, there is a greater probability of the median being greater than the 0.45 threshold, but still a relatively large chance that we are below the threshold. If we increase the number of samples from $n = 50\ to\ n = 500$, we find that the probability of being below the threshold is greatly reduced:

```
Threshold value =  0.45
Probability that percentile < threshold:  0.01678333
```

Obviously this is dependent on our random sample, but the result is consistent with expectation.

Example 2: We take a sample of a uniformly distributed random variable with $n = 50$ between the values of 0 and 1. If calculate the probability that the 95th percentile is less than 0.95 we find:

```
Threshold value =  0.95
Probability that percentile < threshold:  0.5176459
```

If we increase the number of samples from $n = 50\ to\ n = 500$:

```
Threshold value =  0.95
Probability that percentile < threshold:  0.4801408
```

Again, the results depend on the distribution of our random uniform sample, but the results are consistent with expectation.

One limitation of this method is that the probability values will be more sensitive (and therefore, less reliable) when the percentile point estimate (numeric current attribute state) is close to the threshold value (numeric TAS), but more reliable when these two values are more different. However, this is likely to be the case with any statistical method to compare numeric current attribute state and TAS.

For example, compare Example 2 above, where the percentile and threshold value were very close and therefore we obtained an uncertain result (probability close to 0.5), to the following example. We take a sample of a uniformly distributed random variable with $n = 50$ between the values of 0 and 1. If calculate the probability that the 95th percentile is less than 0.5 we find:

```
Threshold value =  0.5
Probability that percentile < threshold:  < 0.005
```

This result is very repeatable for any uniform random sample, and gives good confidence that the 95th percentile of the data is not less than the threshold value of 0.5.

Given the sensitivity described above, it may be worth classifying the probability into bins, such as those presented as in Snelder et al. (2021) and reproduced as Table B-1 below.

Table B-1: Probability that a calculated percentile is below a pre-defined threshold divided into bins of confidence. The number of categories could be reduced (e.g., in line with the five presented in Table 3-1 and MfE (2018)).

Probability that percentile is less than a threshold value	Confidence rating
>0.99	Virtually certain
0.95 – 0.99	Extremely likely
0.9 – 0.95	Very likely
0.67 – 0.9	Likely
0.33 – 0.67	As likely as not
0.1 - 0.33	Unlikely
0.05 – 0.1	Very unlikely
0.01 – 0.05	Extremely unlikely
< 0.01	Exceptionally unlikely

Appendix C R code for confidence and tolerance interval calculations

The following code was provided by Horizons Regional Council.

```
####################################################################################
#
# Example code to:
#     a) calculate confidence interval for a percentile.
#     b) compare the state of an attribute to a threshold value
#
####################################################################################

rm(list = ls())                 # clear the memory
########################################################
alpha_level = (1-0.9)  #1-alpha, for alpha Confidence Interval
percentile = 0.95      # Desired percentile (between 0 and 1)

####################################################################################
####################################################################################
####################################################################################
#Wilson CI function for percentiles
binom.CI <- function(events, #events = outcomes
              trials, #number of individuals, test, etc
              alpha = 0.05){
 n <- trials
 x <- events
 p.hat <- x/n
 # Calculate upper and lower limit
 upper.lim <- (p.hat +
         (qnorm(1-(alpha/2))^2/(2*n)) +
         qnorm(1-(alpha/2)) * sqrt(((p.hat*(1-p.hat))/n) +
                   (qnorm(1-(alpha/2))^2/(4*n^2))))/(1 + (qnorm(1-(alpha/2))^2/(n)))
 lower.lim <- (p.hat +
         (qnorm(alpha/2)^2/(2*n)) +
         qnorm(alpha/2) * sqrt(((p.hat*(1-p.hat))/n) +
                  (qnorm(alpha/2)^2/(4*n^2))))/(1 + (qnorm(alpha/2)^2/(n)))
 # Modification for probabilities close to boundaries
 if ((n <= 50 & x %in% c(1, 2)) | (n >= 51 & n <= 100 & x %in% c(1:3))) {
  lower.lim <- 0.5 * qchisq(alpha, 2 * x)/n
 }
 if ((n <= 50 & x %in% c(n - 1, n - 2)) | (n >= 51 & n <= 100 & x %in% c(n - (1:3)))) {
  upper.lim <- 1 - 0.5 * qchisq(alpha, 2 * (n - x))/n
 }
 out <- c(lower.lim,upper.lim)
 return(out)
}
####################################################################################
####################################################################################
####################################################################################
#function to calculate Hazen percentile
```

```
hazen_percentile <- function(x, #data values
                             percentile){
  hz_pc = round(quantile(x, probs = percentile, type = 5),5)
  return(hz_pc)
}
#################################################################################
#################################################################################
#################################################################################
percentile_ranges <- function(x,
                              percentile){
  confidence_levels_to_check = c(0.55, 0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9, 0.95, 0.975, 0.99)
  confidence_level = c()
  one_sided_confidence_level = c()
  lower_percentile = c()
  upper_percentile = c()
  lower_value = c()
  upper_value = c()

  #append case with no confidence interval
  confidence_level = append(confidence_level,0.5)
  one_sided_confidence_level = append(one_sided_confidence_level,0.5)
  lower_percentile = append(lower_percentile,percentile)
  upper_percentile = append(upper_percentile,percentile)
  lower_value = append(lower_value,unname(hazen_percentile(x,percentile)))
  upper_value = append(upper_value,unname(hazen_percentile(x,percentile)))

  #iterate through confidence levels
  for (conf_level_i in confidence_levels_to_check){
    confidence_level = append(confidence_level, conf_level_i)
    one_sided_confidence_level = append(one_sided_confidence_level, (1-
conf_level_i)/2+conf_level_i)
    cis = (binom.CI(percentile*length(x), length(x), 1-conf_level_i))
    lower_percentile = append(lower_percentile,cis[1])
    upper_percentile = append(upper_percentile,cis[2])
    lower_value = append(lower_value,unname(hazen_percentile(x,cis[1])))
    upper_value = append(upper_value,unname(hazen_percentile(x,cis[2])))
    }
  df <- data.frame(confidence_level, one_sided_confidence_level, lower_percentile,
upper_percentile, lower_value, upper_value)
  return(df)
}
#################################################################################
#################################################################################
#################################################################################
find_threshold_percentage <- function(x, threshold, percentile){
  #Returns the probability that the percentile is less than the threshold value
  df = percentile_ranges(x,percentile)
  haz_pc= unname(hazen_percentile(x,percentile))

  if (haz_pc == threshold){
    cat('Probability that percentile < threshold:  0.5')
    return(0.5)
```

```
  }else if (haz_pc > threshold) {
    if (threshold<min(df$lower_value)){
      cat('Probability that percentile < threshold:  <', 1 - max(df$one_sided_confidence_level))
      return(max(df$one_sided_confidence_level))
    }
    confidence_column = df$lower_value
    pred <- approx(y=df$one_sided_confidence_level, x=confidence_column , xout=threshold, rule=2)
    cat('Probability that percentile < threshold: ', 1 - pred$y)
  } else {
    if (threshold>max(df$upper_value)){
      cat('Probability that percentile < threshold: >', max(df$one_sided_confidence_level))
      return(max(df$one_sided_confidence_level))
    }
    confidence_column = df$upper_value
    pred <- approx(y=df$one_sided_confidence_level, x=confidence_column , xout=threshold, rule=2)
    cat('Probability that percentile < threshold: ', pred$y)
  }
  return(pred$y)
}
####################################################################################
####################################################################################
####################################################################################


#  a) calculate confidence interval for a percentile.

####################################################################################
#Generate data
set.seed(1234)
mydata = rnorm(50)  #normal distribution
#mydata = rlnorm(50) #log-normal distribution
mydata = runif(50,0,1) #Uniform distribution between 0 and 1
####################################################################################


ci = (binom.CI(percentile*length(mydata), length(mydata), alpha_level))
cat('The confidence limit for the ',percentile ,'th percentile = (', ci[1],', ', ci[2],') \n')


cat('The confidence limits of the data: (')
cat(hazen_percentile(mydata,ci[1]),', ',hazen_percentile(mydata,ci[2]),')')

cat('\n Hazen point estimate for the percentile: ', hazen_percentile(mydata,percentile), '\n')


#  b) compare the state of an attribute to a threshold value.
threshold_value = 0.9
cat('Threshold value = ', threshold_value, '\n')

df = percentile_ranges(mydata,percentile)

prediction = find_threshold_percentage(mydata, threshold_value, percentile)
```