Document title	Further development of the confidence assessment in HEAT 3.0
Code	4-2
Category	CMNT
Agenda Item	4-Improvements in the HEAT assessment procedure
Submission date	19.5.2021
Submitted by	Germany
Reference	

Background

This document addresses the further development of the confidence assessment in HEAT 3.0 in particular regarding the accuracy aspect. IN-EUTROPHICATION 19-2021 took note that for the accuracy aspect of the confidence assessment in HEAT, the Monte Carlo simulation cannot be implemented in R in the short-term as planned due to existing statistical difficulties related to the NEAT tool of the DEVOTES project from where it should have been transferred to BEAT and HEAT. An alternative calculation for estimating the probability of classification is described in this document including test assessment results of HEAT 3.0.

Action requested

The Meeting is invited to take note of the information in this document, discuss the proposal for including accuracy in the confidence assessment in HEAT 3.0 and agree on further steps in the HEAT assessment procedure.

Introduction

This document provides an overview of the suggested procedure for the accuracy aspect of the confidence assessment in HEAT 3.0. Test assessment results are shown for different indicators in selected areas. Based on the outcome of IN-EUTROPHICATION 19-2021, it was agreed that temporal, spatial and accuracy confidence should be included in the assessment and was therefore summarized in the document on improvements in the HELCOM eutrophication assessment tool (4J-80) for State & Conservation 14-2021. It would be beneficial if methodological confidence could be included in the assessment in the next step to ensure a similar procedure as in the BEAT assessment tool, and thus contribute to harmonisation of assessments.

Accuracy confidence

To consider the accuracy confidence aspect it was suggested to include the Monte Carlo simulation similar to the HELCOM BEAT tool as already proposed in previous documents to include a statistical resampling method to determine the uncertainty or probability of the classification results. It was assumed that the Monte Carlo simulation was already implemented in BEAT and applied in HOLAS II, but this turned out to be not correct and a short-term solution of the statistical difficulties is not possible. However, in order to include the important accuracy aspect in the assessment, an alternative calculation for estimating the probability of classification has been elaborated based on the confidence rating in the eutrophication assessment of OSPAR. It has already been test-implemented in HEAT to investigate the results in more detail.

The accuracy of the indicator result indicates how certain the assessment is in relation to the variability of the data. The accuracy aspect of the confidence assessment is considered by calculating variable confidence level per assessment indicator to estimate the probability or certainty of the classification of being below or above the area-specific thresholds (depending on the response of the indicator to eutrophication) and thus the classification as failing or achieving GES. In contrast to temporal and spatial confidence, the accuracy will be assessed over the entire assessment period and not on annual basis, because it is a matter of estimating the probability of correct classification for the overall result.

The variable confidence level is calculated in the assessment procedure of HEAT based on the observed value (ES), the target value (ET) and the standard error of the respective assessment indicator per assessment area on the basis of the normal distribution function. The calculated confidence level is directly used as the probability of correct classification for good or not good status. The class boundaries for the accuracy confidence are taken directly from the BEAT assessment to ensure a harmonised approach as far as possible and are listed in Table 1 below. In case of missing information on standard deviation, number of observations and standard error, no calculation of variable confidence levels and thus no quantitative accuracy estimates will be possible. Alternatively, a qualitative estimate based on expert judgement for the respective indicator and area can be used.

Table 1: Confidence class boundaries for accuracy confidence (ACC)

Score	Evaluation criteria for accuracy confidence (confidence level of being above or below area-specific assessment level)
HIGH	Assessment result is considered correct with > 90 % probability
MODERATE	Assessment result is considered correct with a probability between 70 % and 90 %
LOW	Assessment result is considered correct with less than 70 % probability

Test assessment results have been investigated for several indicators and both assessment periods of HOLAS II and the upcoming HOLAS III. The majority of results showed high accuracy of classifications into good or not-good status. In some areas the accuracy was moderate, either due to oberserved values close to target values, relatively high standard errors and/or a lower number of observations. Selected examples of assessment results for different areas and indicators in both assessment periods are shown in Table 2.

Table 2: Test assessment results of accuracy confidence (ACC) for different indicators in selected example areas (ET: target value, ES: observed value, SD: standard deviation, n: number of observations, SE: standard error)

Period	HELCOM ID	Assessment unit	Indicator	ET	ES	SD	n	SE	ACC	ACC class	status
2011-2016	SEA-008	Gulf of Gdansk	DIN	4.2	4.59	2.47	32	0.44	81.8%	M	not good
2016-2021	SEA-008	Gulf of Gdansk	DIN	4.2	7.36	2.24	23	0.47	100.0%	H	not good
2011-2016	SEA-004	Kiel Bay	Secchi	7.4	6.93	0.49	197	0.04	100.0%	H	not good
2016-2021	SEA-004	Kiel Bay	Secchi	7.4	7.38	0.60	110	0.07	63.8%	L	not good
2011-2016	SEA-004	Kiel Bay	DIN	5.5	5.91	0.86	88	0.09	99.9%	H	not good
2016-2021	SEA-004	Kiel Bay	DIN	5.5	4.50	2.13	35	0.36	99.7%	H	good
2011-2016	SEA-001	Kattegat	Chl-a	1.5	0.98	0.25	300	0.01	100.0%	H	good
2016-2021	SEA-001	Kattegat	Chl-a	1.5	1.18	0.21	204	0.01	100.0%	H	good

The selected examples reflect different situations in the sub-basins and also differences and developments between assessment periods. Accuracy confidence for DIN in the Gulf of Gdansk was moderate for classification as not good status in the period 2011-2016. The ES value was above the threshold, but still within the range of the threshold due to the relatively high standard error. This is reflected in the moderate classification below 90% probability. In the upcoming assessment period 2016-2021, the ES value is significantly higher (current status of ICES data base, more data expected), and although the number of observations is lower, the standard error is in the same range as in the previous period. The resulting accuracy confidence would be high, meaning a certainty of 100% of being in not good status.

The assessment of Secchi depth in the Kiel Bay resulted in not good status in both assessment periods, but with different accuracy confidence. The observed value in the period 2011-2016 was clearly below the threshold, the number of observations was high and the standard error low, thus the probability for classification of not good status was high (100%). Measurements in the period 2016-2021 show an increase of Secchi depth and the ES value is close to the target value with a lower number of observations compared to 2011-2016 (reflected in lower temporal and spatial confidence), but still a relatively low standard error. The resulting accuracy confidence would be low for classification of being in not good status (63.8%), as the measurements show an improved situation for Secchi depth and are very close to the threshold value.

Accuracy confidence for DIN in the Kiel Bay was high in the test assessment for both assessment periods, but reflected different status assessments. The observed value in the period 2011-2016 was above the threshold, the number of observations was high and the standard error low. This resulted in a probability of 99.9% for being in not good status. In the period 2016-2021, the measurements show decreased concentrations and the ES value is clearly below the threshold. Even though the number of observations has decreased significantly and the standard error is much higher than 2011-2016, the resulting accuracy confidence would be high for being in good status. The reduced monitoring is reflected in lower confidence scores for both temporal and spatial confidence compared to the period 2011-2016.

The assessment of chlorophyll-a in the Kattegat showed good status in both assessment periods with ES values clearly below the threshold. The number of observations was high and the standard error quite low, so that the probability of classification for good status was 100%.

This clearly shows that the different confidence aspects reflect very well the respective conditions based on measured values in relation to the threshold as well as the different variability of the indicators and monitoring efforts in the different assessment areas. The results of the different confidence aspects should be presented both separately and as an overall assessment using maps to illustrate these differences. Work is currently underway to automatically generate confidence maps in the HEAT tool.

Methodological confidence

The aspect of methodological confidence considers the quality of the monitoring methodology. In contrast to temporal, spatial and accuracy aspects, methodological confidence is not a data-driven assessment in HEAT. Instead, predefined confidence values are used for the different indicators based on the level of monitoring conducted according to HELCOM Guidelines and on the quality assurance. Following the confidence classes applied in the BEAT tool for the methodological aspect, high confidence is assigned if the monitoring has been conducted according to HELCOM Guidelines (for indicators where these are available) and the data is quality assured according to HELCOM or other internationally accepted guidelines. Moderate confidence is assigned if the monitoring has been conducted only partly according to HELCOM Guidelines and/or the data originates from mixed sources and is partly quality assured according to HELCOM or other international standards and/or the data is quality assured, but according to local standards. If monitoring has not been conducted according to HELCOM Guidelines or the data has not been quality assured, the methodological confidence is considered low.

Following the confidence classes for the methodological aspect applied in the BEAT tool, the indicators DIN, DIP, TN, TP, Secchi and Benthic Quality Index would be assigned a high confidence rating, while the indicators with mixed data sources such as chlorophyll-a, cyanobacterial blooms and oxygen debt would be given a moderate rating.

It should be discussed if the methodological confidence could be included in the HEAT procedure with predefined confidence values according to the confidence classes as used in the BEAT tool or if further adaptations are necessary for eutrophication indicators and the different data types used in the assessments.