ICS 07.060
CCS A45

HY

中 华 人 民 共 和 国 海 洋 行 业 标 准

XX/T XXXXX—XXXX

海洋温盐数据排重技术规范

Technical specification for eliminating duplicates of oceanographic temperature and salinity data

点击此处添加与国际标准一致性程度的标识

（报批稿）

（本稿完成日期：2022 年 7 月 10 日）

XXXX－XX－XX 发布　　XXXX－XX－XX 实施

中华人民共和国自然资源部　发布

目　次

前　言

本文件按照 GB/T 1.1-2020《标准化工作导则 第 1 部分：标准化文件的结构和起草规则》的规定起草。

请注意本文件的某些内容可能涉及专利。本文件的发布机构不承担识别专利的责任。

本文件由中华人民共和国自然资源部提出。

本文件由全国海洋标准化技术委员会（SAC/TC283）归口。

本文件起草单位：国家海洋信息中心。

本文件主要起草人：纪风颖、刘玉龙、徐珊珊、董明媚、岳心阳、骆敬新。

海洋温盐数据排重技术规范

1 范围

本文件规定了海洋常用观测仪器获取的温盐数据排重的原则、流程、方法和参数。
本文件适用于11种仪器所获得的海洋温盐数据的排重工作。

2 规范性引用文件

本文件没有规范性引用文件。

3 术语和定义

下列术语和定义适用于本文件。

3.1

关键信息项 key item

表征温盐数据的数值、精度以及观测仪器等字段。
注：包括观测仪器、观测时间、观测经纬度、观测层深、温度和盐度。

3.2

辅助信息项 assistant item

用于说明温盐数据来源的背景信息。
注：包括调查国家、调查机构、调查项目、整编项目、数据集名称等。

3.3

重复数据 duplicate data

由于重复接收、收集或存储所造成的关键信息项完全相同的数据或满足从属关系的数据。

3.4

准重复数据 near duplicate data

观测时间和经纬度相同，且水下观测数据90%以上相同的测站数据。

4 海洋温盐重复数据的排除原则

海洋温盐数据排重工作应确保最终成果数据来源真实、信息完整、数据可靠和结果唯一，具体原则包括以下内容。

a) 来源真实:应保留从原始调查者获取的一手数据，剔除再收集数据。
b) 信息完整：
   1）应保留资料内容、类型以及元数据信息完整的数据，剔除片段数据；
   2）应优先保留主体观测仪器的资料，删除或标识辅助的比测资料。
c) 数据可靠：应保留观测数据精度与仪器性能相匹配的数据，剔除人工修改/插值及多次转存精度变化的数据。
d) 结果唯一：排重工作完成后，应保证成果数据中不再有重复记录。

5 海洋温盐数据排重流程

5.1 计算机判定标识海洋温盐重复数据

5.1.1 计算机判定海洋温盐重复数据的关键信息项

根据现有温盐数据的组成和存储方式，将海洋温盐重复数据判定的关键信息项设定为观测仪器、观测时间、观测经纬度、观测层深、温度和盐度。

5.1.2 海洋温盐数据关键信息项相同的判定阈值

对目前应用广泛（获取数据最多）的11种海洋温盐观测仪器：颠倒温度计、温盐深仪（Conductivity Temperature Depth, 简称CTD）、机械式温深仪（Mechanical Bathythermograph,简称MBT）、抛弃式温深仪（Expendable Bathythermograph,简称XBT）、海表面温盐记录仪、海洋游泳动物携带温盐深仪、锚定浮/潜标、剖面浮标（简称Argo）、漂流浮标（本文件中漂流浮标包括漂流温度链或CTD链）、拖曳式温盐深仪（Undulating Conductivity Temperature Depth,简称UCTD）以及水下滑翔机（Glider），依据其传感器精度、观测方式和数据传输方式，判定海洋温盐数据是否重复的阈值，即关键信息项相同的阈值包括如下内容。

a) 本文件推荐测站位置相同的阈值为 10m。具体使用过程中，可根据卫星定位系统精度，以及温盐数据获取过程中位置的变动范围进行调整。
b) 测站时间相同阈值包括如下内容。
   1) 对于人工下放观测的颠倒温度计、CTD、MBT 和 XBT，观测时间相同的阈值为完成一个测站观测所需的最少时间；
   2) 对于自动观测的 Argo 浮标、漂流浮标、Glider、和 UCTD，观测时间相同阈值为完成一次观测所需时间的最小值，以 Glider 为例，其完成一次观测基本为 3h～9h，则该仪器观测时间相同的阈值为 3h；
   3) 对于海洋游泳动物携带温盐深仪，本文件推荐观测时间相同阈值为 1min，实际工作中可根据动物下潜和上浮时间间隔进行调整；
   4) 对于志愿船携带的海表面温盐记录仪，本文件推荐该仪器观测时间相同的阈值为 10s，实际工作中，可根据仪器观测频率和测站的实际时间间隔调整该参数。
c) 本文件推荐观测层深相同的阈值为 1m。实际工作中可根据观测仪器所携带的压力传感器的精度进行调整。
d) 本文件推荐水下温盐数据相同的阈值包括如下内容。对于观测主体为 CTD 的下放式 CTD、XCTD、UCTD、漂流浮标、Argo、Glider、海洋游泳动物携带温盐深仪，观测数据相同的标准为同层观测数据温度差异不大于 0.01℃，盐度差异不大于 0.01。对于观测主体为 BT 的 MBT、XBT 和海面浮子，综合其传感器精度，同层温度相同的判定阈值为 0.1℃。

以上仪器获取温盐数据是否重复的判定阈值见表1。

表1　不同仪器获取的温盐数据重复的阈值

观测仪器	测站时间间隔	测站距离 m	同层温度差异 ℃	同层盐度差异 PSS-78
CTD	2h	10	0.01	0.01
UCTD	1h	10	0.01	0.01
Argo	6 h	10	0.01	0.01
Glider	3h	10	0.01	0.01
漂流浮标	1h	10	0.01	0.01
颠倒温度计	1h	10	0.01	0.01
锚定浮/潜标	5 min	10	0.1	0.1
海洋游泳动物携带温盐深仪	1min	10	0.1	0.1
海表面温盐记录仪	1min	10	0.1	0.1
XBT	10min	10	0.1	无
MBT	2h	10	0.1	无

实际温盐数据排重工作中可以根据传感器精度、观测方式和站位水深等信息，并结合排重结果，对表1中各种阈值进行适度调整。

对于表1中未列入的观测仪器，温盐重复数据判定阈值可依据传感器精度和观测方式进行设定。

5.1.3　海洋温盐重复数据的计算机判定算法

5.1.3.1　测站时间和位置重复的判定

任意测站A站和B站，当二者的距离不大于10m，而且观测时间差异小于该仪器观测时间相同的阈值，则认为A和B为可能重复测站，进行下一步判断，否则二者非重复数据，不再进行以下判断过程。

5.1.3.2　水下温盐数据重复的判定

对于 5.1.3.1 中判断的可能重复测站 A 站和 B 站，为描述方便，命名总层次数较多的测站为 A 站，总层次数据较少的测站为 B 站，即 A 的层次数大于等于 B 的层次数，并按照 5.1.2 中的 c）取出 A 和 B 中相同层深的数据，分别组成新的子集 C（C∈A）和 D(D∈B)。

判断 A 站与 B 站水下温盐数据是否重复的算法具体如下：

a)　当 D 层次个数小于 B 总层次数的 90%，则 A、B 两个站水下温盐数据不重复；

b)　当 C 和 D 所有温盐数据满足 5.1.2 的温盐重复数据的参数，则对 B 站水下温盐数据进行重复数据的标识；

c)　当 C 和 D 同层温盐数据不满足 5.1.2 的温盐重复数据的参数，则 A、B 两个站水下温盐数据不重复。

5.1.4　海洋温盐重复数据的标识

根据海洋温盐数据是否重复的计算机判定结果，对相应的测站进行标识，为后期人工审核和确认提供依据。具体标识方法如下：

a)　对于未出现重复数据的站点，站点重复数据标识符均设定为“ 0 ”；

b) 对于 A、B、C 等多个测站，所有关键信息项相同，则将这些站点数据归为同一个数据组，站点重复数据标识符均设定为“ 1 ”；
c) 对于 A、B、C 等多个测站，时间地点相同，但是 B、C 等站点的观测数据为 A 的子集，则将 A、B、C 等站点归为同一个数据组，站点重复数据标识符分别设定为“ 1 ”、“2”、“2”等。

5.2 海洋温盐重复数据人工审核和确认

5.2.1 海洋温盐重复数据判定的辅助信息项

辅助信息项包括调查国家、调查机构、调查项目、调查航次、调查船、首席专家等信息。

5.2.2 海洋温盐重复数据人工审核和标识

计算机自动判断和标识重复数据和准重复数据后，需人工对标识符进行审核。基于关键信息项的差异情况，结合相应的辅助信息项，对温盐重复数据进行认可或修正，具体操作如下；

a) 对于关键信息项完全相同的数据组，对调查资料进行溯源，明确资料的来源：调查项目、首席专家，对保留原始收集者/负责人获取的观测数的站点重复数据标识符设定为“1”，其它数据的标识符设定为“2”；
b) 对于满足从属关系的重复数据组，若为卫星等自动重复接收同一个站点/仪器的数据，选取数据最为完整的数据，将其站点重复数据标识符设定为“1”，其余相同或片段数据的重复数据标识符设定为“2”；
c) 对于满足从属关系的重复数据组，若为原始数据和后期插值/抽稀数据共存情况，则选取数据最为完整的原始数据，将其站点重复数据标识符设定为“1”，其余相同或片段数据的重复数据标识符设定为“2”；
d) 对于关键信息项完全相同的多来源国际数据的重复数据组，将最初观测计划/收集者获取的观测数据的站点重复数据标识符设定为“1”，其余相同或片段数据的重复数据标识符设定为“2”；
e) 对于收集到来源/观测计划/调查人等不明确的观测数据重复数据组，人工审核其数据精度和辅助信息项，将精度更为接近调查仪器精度、信息更全面的站点重复数据标识符设定为“1”，其余相同或片段数据的重复数据标识符设定为“2”。

5.3 海洋温盐重复数据的剔除

计算机剔除重复数据标识符为“2”的测站数据。

参 考 文 献

[1] Boyer, T.P., J.I. Antonov, O.K. Baranova, C. Coleman, H.E. Garcia, A. Grodsky, D.R. Johnson, R.A. Locarnini, A.V. Mishonov, T.D. O'Brien, C.R. Paver, J.R. Reagan, D. Seidov, I.V. Smolyar, M.M. Zweng, 2013, World Ocean Database 2013. Sydney Levitus, Ed.; Alexey Mishonov, Technical Ed.; NOAA Atlas NESDIS 72, 209 pp.

[2] GTSPP Real-Time Quality Control Manual, First Revised Edition. UNESCO-IOC 2010. (IOC Manuals and Guides No. 22, Revised Edition.) (IOC/2010/MG/22Rev.)

[3] Annie Wong, Robert Keeley, Thierry Carval and the Argo Data Management Team (2020). Argo Quality Control Manual for CTD and Trajectory Data. http://dx.doi.org/10.13155/33951

[4] 纪风颖等，WOD 与 Argo 数据集的排重方法与软件实现，中国海洋大学学报，2015.8.1，45(8)：121~127