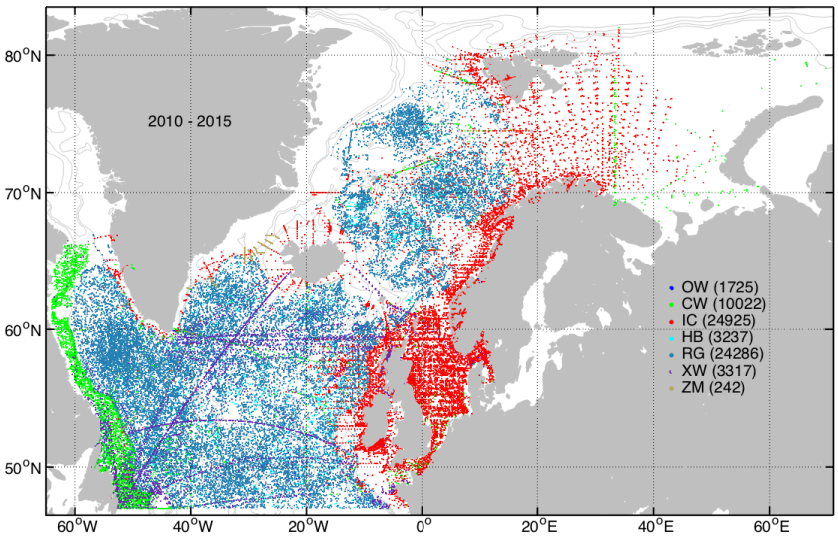
The North Atlantic and Nordic Seas hydrography collection

J. Even Ø. Nilsen

Nansen Environmental and Remote Sensing Center, Bergen, Norway

Bjerknes Centre for Climate Research, Bergen, Norway



NERSC Technical Report no. 372.

Bjerknes Centre for Climate Research

NERSC

Contents

Chapter 1

Introduction

This report describes a collection of hydrographic (temperature and salinity) profiles and surface samples compiled in order to support several projects at the Nansen- and Bjerknes centres involving the North Atlantic and Nordic Seas. The region covered includes the subpolar North Atlantic Ocean (north of 47°N; Newfoundland–Brest), the Nordic Seas, and the Barents Sea (i.e., to 83°N and 70°E). The time span is 1900–2015. The collection is called the North Atlantic and Nordic Seas hydrography collection (NANSHY).

The relatively cold and fresh subpolar gyre and the northward flow of warm and saline waters characterize the region south of the Greenland–Scotland Ridge. In the southeastern part of the Nordic Seas warm and saline water from the North Atlantic flow northward while in the western part of the Nordic Seas cold and relatively fresh water flow southwards. The interaction of water masses and their mixing product is essential for the climate and living conditions in these areas.

The purpose of this collection is not to deliver any kind of gridded product or extensively quality checked and user supported database, but merely to provide a collection of the most useful single profiles and samples, compiled from several databases and sources. It is not intended to be *the* database for the region, neither wrt. quality (see Section 2.5) nor public access (see Section 3.2). The treatment of data involves, however, checking for duplicates, elimination of gappy profiles, and other simple sifting methods.

The work and motivation itself stems from the data collection done in the Project NISE (Norwegian Iceland Seas Experiment Nilsen et al., 2008), but all downloads and code has been remade. The NISE-dataset is among the sources used.

Contact

Jan Even Øie Nilsen, Nansen Environmental and Remote Sensing Center, Thormøhlensgt. 47, N-5006 Bergen, Norway, email: even@nersc.no.

Chapter 2

The Collection

2.1 Data Sources

2.1.1 Platforms

The hydrographic variables included in this collection are *temperature*, and *salinity*. The measurement of these variables are either done by collecting water samples in (Nansen/Niskin) bottles for later analysis and reading reversible thermometers, or electronically using a CTD (Conductivity Temperature Depth) sonde. In the latter, depth is given by the pressure and from the three variables the salinity can be calculated. The CTD is a relatively modern instrument, but bottle sampling are still used especially for calibration of the electronic data.

The concept "station" refers to a geographical position and time where and when hydrographic measurements are made. Each measurement made at a given station is referred to as a "sample", and a vertical sequence of samples is termed a "profile". A horizontal sequence of profiles constitute a "section".

Another type of instrument used in the collection of the data used here, is the float. These are automated CTD-sondes floating at a pre-set depth or density level, surfacing at regular intervals, taking hydrographic profiles during their ascent. These profiles are in this report also referred to as "stations".

Expendable bathythermograp (XBT) data is also included.

2.1.2 Contributors

The collection is built by gathering the following public databases and other contributions:

WOD13 CTD, bottle (**OSD**), and **XBT** data from the World Ocean Database 2013, downloaded in 2015 from *www.nodc.noaa.gov/OC5/SELECT/dbsearch/dbsearch.html*.

ICES CDT and bottle data data from the data centre of the International Council for the Exploration of the Sea, dowloaded in late 2015 from *www.ices.dk*.

HYDROBASE3 CTD, bottle, and float data from Hydrobase3 at Woods Hole Oceanographic Institution, downloaded in 2013 from *www.whoi.edu/science/PO/hydrobase/php/index.php*.

ARGO float data downloaded in 2015 from *www.argodatamgt.org/Access−to−data/Argo−data−selection*.

NISE data provided especially to the Norwegian Iceland Seas Experiment (Nilsen et al., 2008) by the partners,

FFL Faroese Fisheries Laboratory,

MRI Marine Research Institute, Iceland,

IMR Institute of Marine Research, Norway,

GFI Geophysical Institute, University of Bergen, Norway,

AARI Arctic and Antarctic Research Institute, Russia (via partner NERSC),

as well as some **WOCE** profiles. The latest data ingestion to the NISE base was made in 2009.

ZMAW data provided later by Centre for Marine and Atmospheric Sciences during exchange with NISE data.

The number of stations and samples from different sources are listed in Table 2.1.

Table 2.1: Data contributors, data amounts after duplicate selection, and source numbers or cruise labels for identification in the database. Here 'profiles' are stations deeper than 25 m, and shorter profiles or surface samples are called 'shallow stations'.

ID	Subset	Label	Source	Samples	All stations	Profiles	Shallow stations
1	OSD	OW	WOD13	9 071 439	520 251	404 959	115292
2	CTD	CW	WOD13	1 838 245	68 227	60 960	7267
3		IC	ICES	7 754 572	448 106	364 853	83253
4		HB	HYDROBASE3	7 420 674	269 797	226 498	43299
5	FFL	FA	NISE	807 231	6 885	6 882	3
6	MRI	IS	NISE	165 760	4 667	4 643	24
7	IMR	NO	NISE	62 123	1 793	1 746	47
8	GFI	GF	NISE	88 290	2 237	2 230	7
9	WOCE	WO	NISE	1 014	57	57	0
10		RG	ARGO	3 320 640	39 326	38 145	1181
11	XBT	XW	WOD13	3 693 571	130 864	127 010	3854
12		ZM	ZMAW	120 867	1 784	1 782	2
13	AARI	aa	NISE	207 856	12 666	12 660	6
Total				34 552 282	1 506 660	1 252 425	254 235

2.2 Workflow

2.2.1 Import

Data from each source is separately imported to separate sets in Ocean Data View (ODV; Schlitzer, 2006), either directly or via own reading routines for the different formats. In ODV simple geographical and timespan cropping, first order duplicate checks, and sort and condence is performed. All within the same source.

The software ODV works very well for combining, systemising, and first check of data from different files, however it is chosen to do all further treatment of data in MATLAB, because of the vast amounts and need for total control of the subsequent processing. Hence, the datasets are exported to ASCII files to be read by MATLAB routines. All coding and further processing is done using MATLAB.

2.2.2 Decimation

Profiles from each source are decimated. The maximum density allowed is 5 m in the upper 100 m, 10 m to 500 m, 50 m to 2000 m, and 100 m for further depths. Decimation is done in such a way that samples closest to the grid are kept. As a general rule, no data is altered during processing, only selection is done.

Furthermore, the decimation routine eliminates all single samples deeper than 5 m. Note that there were also empty profiles in the data exported from ODV, which are also eliminated during decimation.

After a re-import, sorting and condensing in ODV, and export (Section 2.2.1), the datasets from each source are read into MATLAB-files (Section 2.3).

2.2.3 Combining into pentads

In order to speed up processing, each source dataset is split into pentads, from 1900–1905 to 2010–2015. For the same reason, all further treatment and storage is done in these pentads.

Then datasets from each pentad from each source, are combined in sets for each pentad.

2.2.4 Duplicate removal

Given that the sources are databases covering the same time and space, there are many duplicates. It is also a fact that profiles have been treated differenty between sources, so the duplicates are not identical. Furthermore, profiles from similar time and place are also considered redundant here. Hence, an advanced duplicate selection routine has been made and used.

First, the duplicates has to be identified. The criteria used here are simple, as follows, and both have to be satisfied

Time Stations in the same half of a day.

Space Stations within 0.01° longitude or latitude distance of each other.

Other metadata, such as cruise and station ID, are not used, as these may be different for the same stations, when coming from different sources. Some databases do not provide time of day, and such stations will be considered to be from the first half of day.

The simple critera also means that there may be actual different stations within a group of 'duplicates' here. This is however not considered problematic, since reduction of data is a priority and a daily resolution is considered more than sufficent for the targeted usage of this collection.

Second, redundant profiles have to be removed. The selection of which duplicate(s) to keep is much more elaborate than the identification. For each group of duplicates, the routine runs through a series of tests, all with the purpose of finding the most useful profile:

Length Profiles of length less than 50% compared to others, are eliminated.

Both variables When there are stations with both temperature (T) salinity (S) profiles, all stations without both are eliminated.

True duplicates All identical profiles but one, are eliminated.

Large gaps Profiles with gaps rendering them useless are eliminated given there is a better profile for that variable, and that the case is not oposite for the other variable.

Similarity If profiles are similar within 10% of both their total temperature and salinity range[1], in at least 80% of their depth range, all but one of them are eliminated, as long as it does not result in loss of depth coverage or severe reduction in sample density.

In summary, the best depth coverage and sample density is sought, while retaining the possibility to calculate density.

Generally there will be an inherent prioritization due to the sorting of sources in the combined dataset. This is deliberate, as there are sources more preferred than others, due to known level of quality control and sharing policies. However, this only kicks in when all else is equal. For instance for true duplicates or very similar stations. The sequence in which sources are sorted is the same as in Table 2.1.

For the single surface values the one sample closest to the mean of all duplicates in the group is selected (minimal RMSE of both temperature and salinity). Remember, no data is altered during processing, only selection is done.

The removal of duplicates resulted in a 40% reduction of the number of stations in the combined dataset.

2.3 Files

The dataset is collected in separate folders for the pentads, in `.mat` files with the following names:

`ny_d` Depth in meters.

`ny_temp, ny_sal` Temperature and salinity.

`ny_t` Time in serial days.

[1] The minimum similarity criteria are 0.01°C and 0.05 for salinity

`ny_lon, ny_lat` Position in degrees.

`ny_BDepth` Bottom depth in meters.

`ny_source` Identification number for source database (see Table 2.1).

`ny_no` Unique number for each profile in the whole set (i.e., pentad, or other subset of files).

`ny_Cruise, ny_Cru, ny_CRUISEn` Cruise label as string, two-character code, and numeric ID[2], respectively.

`ny_Station, ny_STAn` Station label as string and numeric ID[2].

`ny_Type` Sample type as single character.

All variables are single column matlab objects (i.e., N x 1 matrices) with names as last part of filenames, all corresponding to each other point by point. Note that all but the depth, temperature, and salinity, are identical throughout each profile. This may not be storage efficient, but ensures consistency and ease of use. Missing data are represented by NaNs (e.g., when there are temperature samples but no salinity samples).

For extraction of profiles, the following variable is provided:

`ny_stations` Vector of indices to start of profiles in the objects listed above.

Hence, the data can simply be read by MATLAB with *load* and, if needed, profiles can be identified and selected using the station indices.

2.4 Data Coverage

2.4.1 Horizontal Coverage

The region covered includes the subpolar North Atlantic Ocean (north of 47°N; Newfoundland–Brest), the Nordic Seas, and the Barents Sea (i.e., to 83°N and 70°E). Due to low focus and large data concentrations. The Baltic Sea, White Sea and Kara Sea are not included.

Figures 2.1–2.3 show the area coverage in detail. The panels for the different pentads, show the slow, gradual increase in ocean sampling from the early to mid 20th century, the non existense of open ocean scientific activity during wartime, and then the accelleration of sampling through the last half of the century, and the more spatial unform sampling with the advent of ARGO floats at the turn of the century. The history of the other data sources can be seen, but this is however soewhat biased by the duplicate selection (i.e., reflecting the different practices of processing, decimation, etc.). Regions particularly well covered throughout time are the North Sea, the southern Barents Sea, and other near coastal regions in the east. It is only in the second half of the century that there is any coverage worth mentioning in the open ocean.

Near surface samples are shown in the same way in Figures 2.4–2.6. These are separated from the 'proper' profiles, to be able to see features unique to surface samples and deeper profiles separately. Partly due to the strict reduction of redundancy on surface samples, there are not so many shallow stations in this dataset. Other than that, the history of sampling frequency and coverage, is similar.

2.4.2 Vertical Resolution

The collection consists of both CTD-data and bottle data. CTD-data are usually given at 1 m intervals, while bottles are usually taken at coarser depth resolution, often at what is called "standard depths" (see below). This depends on the source dataset, and is further limited by the decimation described in Section 2.2.2.

To give an impression of the different depth resolutions used, the vertical distribution of samples for each contributor is shown in Figure 2.7, in a pentad of data abundance and most of the different contributors, 2000–2005.

2.4.3 Temporal Coverage

The time span for this collection is 1900–2015. The distribution of stations over the years are shown in Figure 2.8. Naturally, data is scarce in the first 50 years, and due to the usual data sharing restrictions also in the latest few years.

2.5 Quality Control

2.5.1 Data Quality

The contributing database providers all have their own routines for quality checking prior to delivery. However, documentation for this is, in addition to being difficult to access, beyond the scope of this report. Users of this collection will have to check online documentation or contact the relevant institutions See Sections 2.1.2 and 3.3 for details.

For this collection no thorough quality assurance have been performed, due to limited resources and the large size of the collection. However, a simple outlier test and visual inspection have been done, in order to eliminate clearly erroneous data, using the following ranges:

Depth 0–7000 m

Temperature -2°C–40°C

Salinity 1–40

The lower salinity limit have removed some near-shore data, but these are few and outside the focus of the target projects. A further, but still coarse, visual inspection in T-S space have been performed to eliminate more outliers. Closer inspection for specific regions must be performed by the user.

All final quality checking prior to publication, is ultimately the responsibility of the user!

[2]Cruise and station labels should be used together with source ID, to ensure uniqueness.

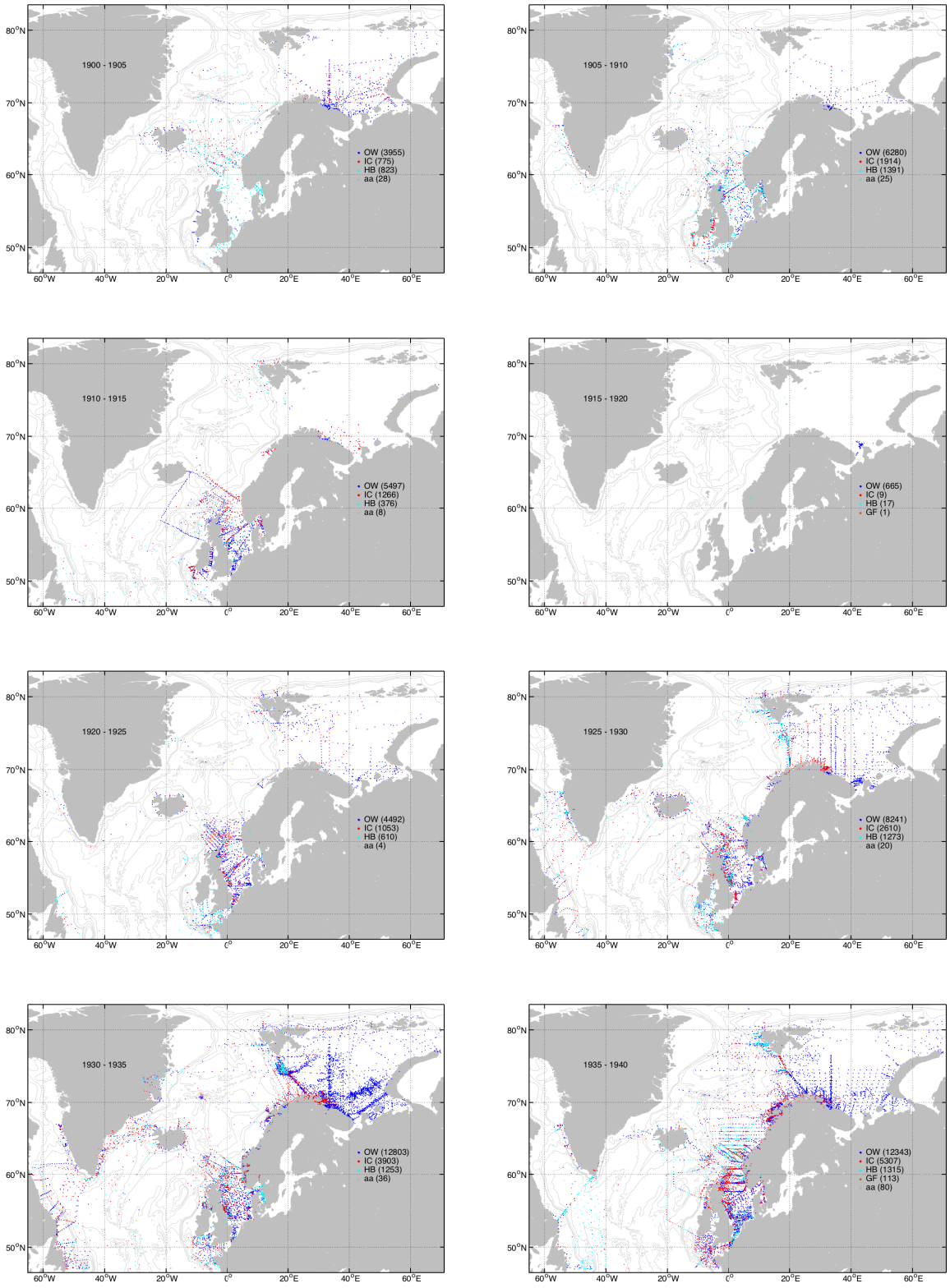


Figure 2.1: Profiles (d_{max} $\geq$25 m) in the dataset, by pentad (panels). Colours indicate which source, and number of stations per source are indicated in the legend.

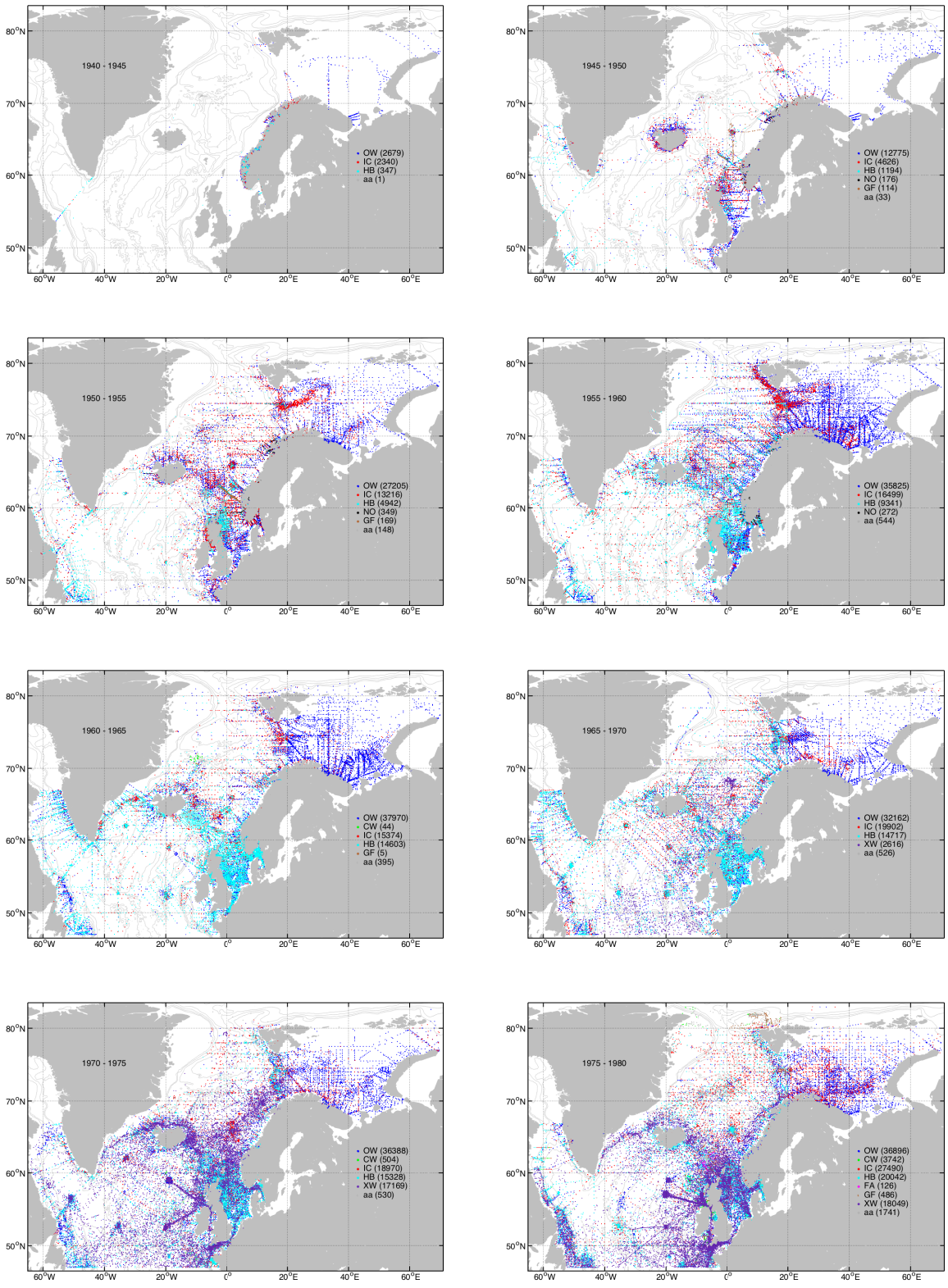


Figure 2.2: Profiles in the dataset by pentad (contd.).

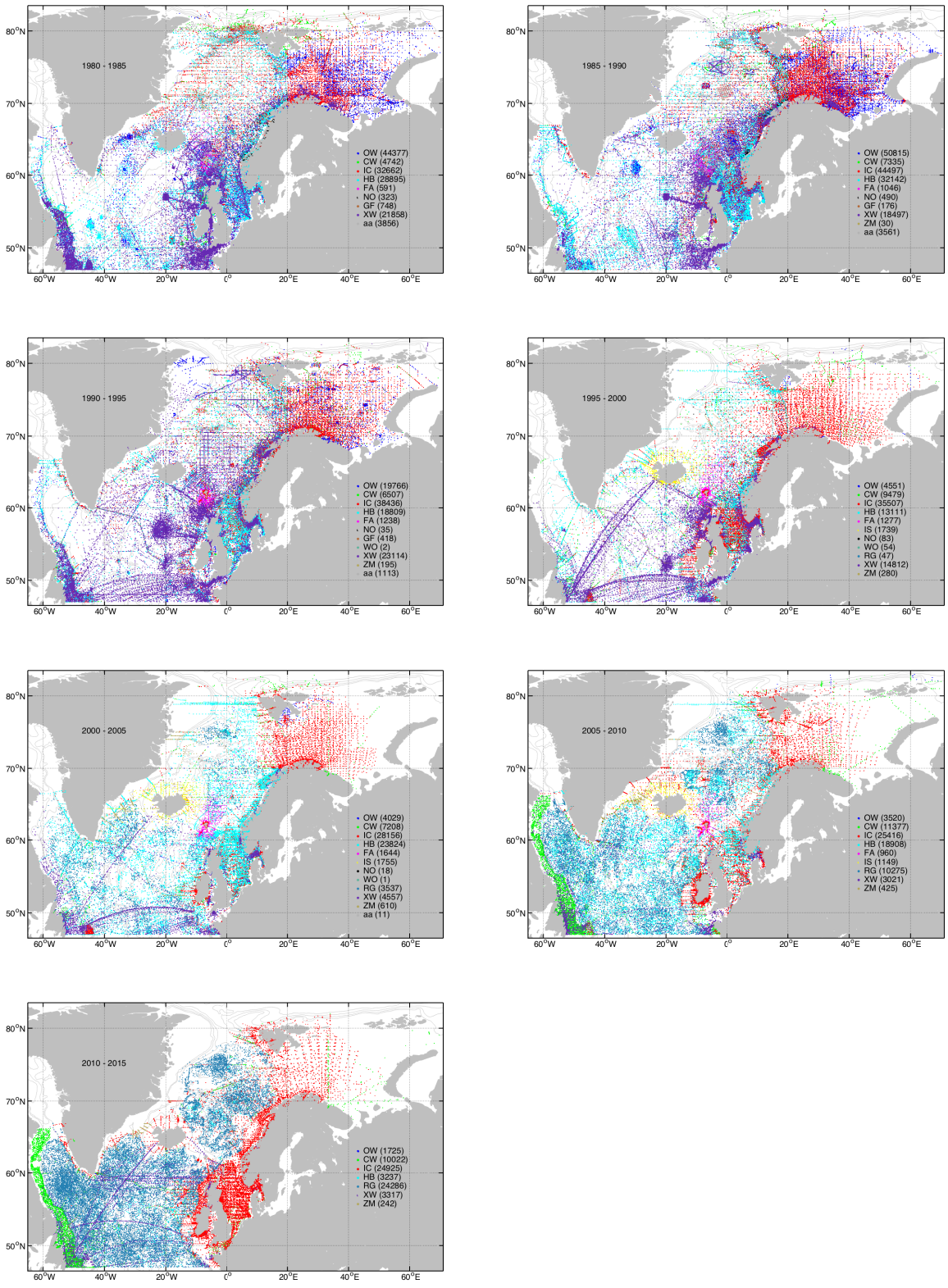


Figure 2.3: Profiles in the dataset by pentad (contd).

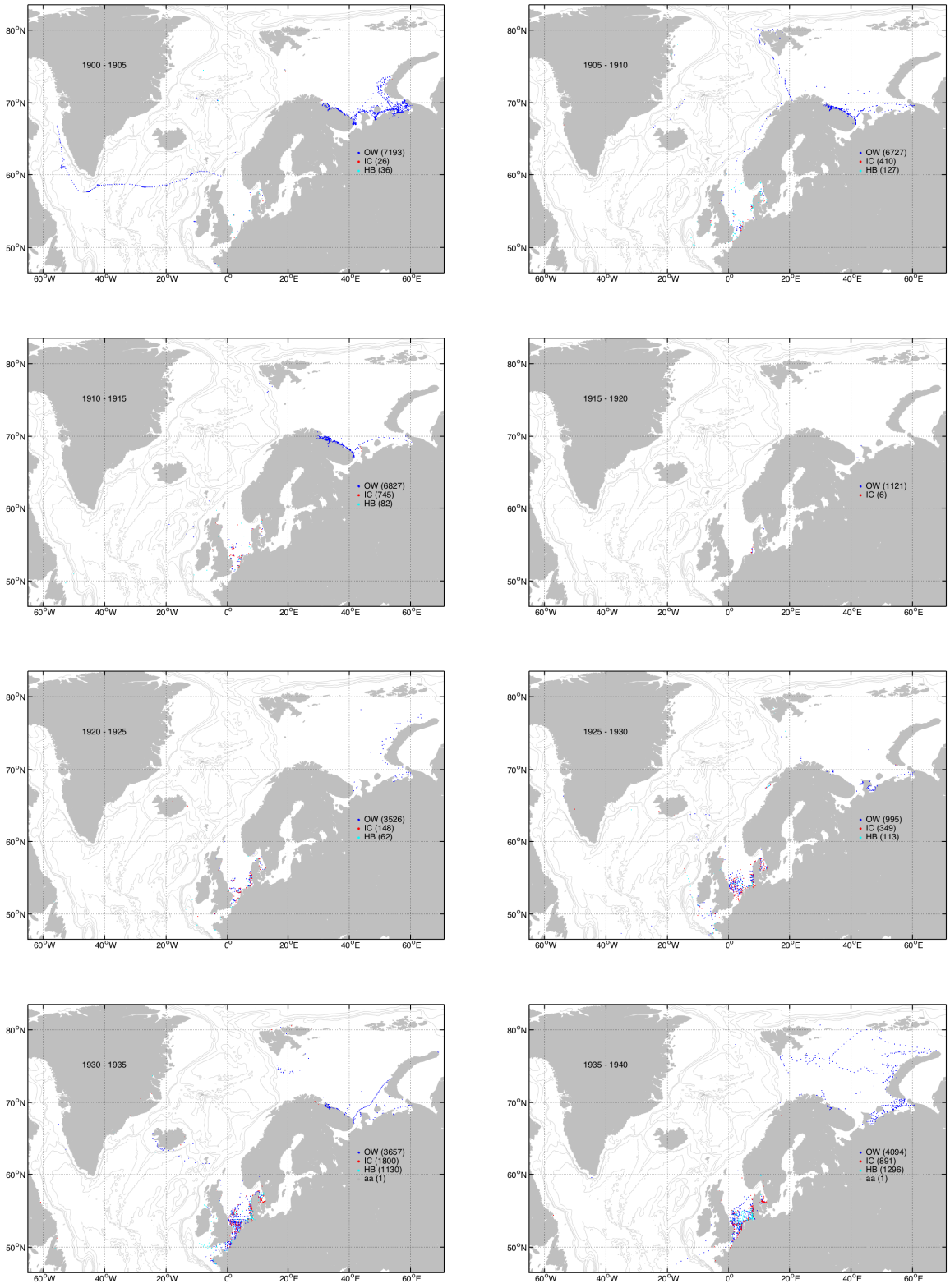


Figure 2.4: Shallow stations (d_{max} <25 m) in the dataset, by pentad (panels). Colours indicate which source, and number of stations per source are indicated in the legend.

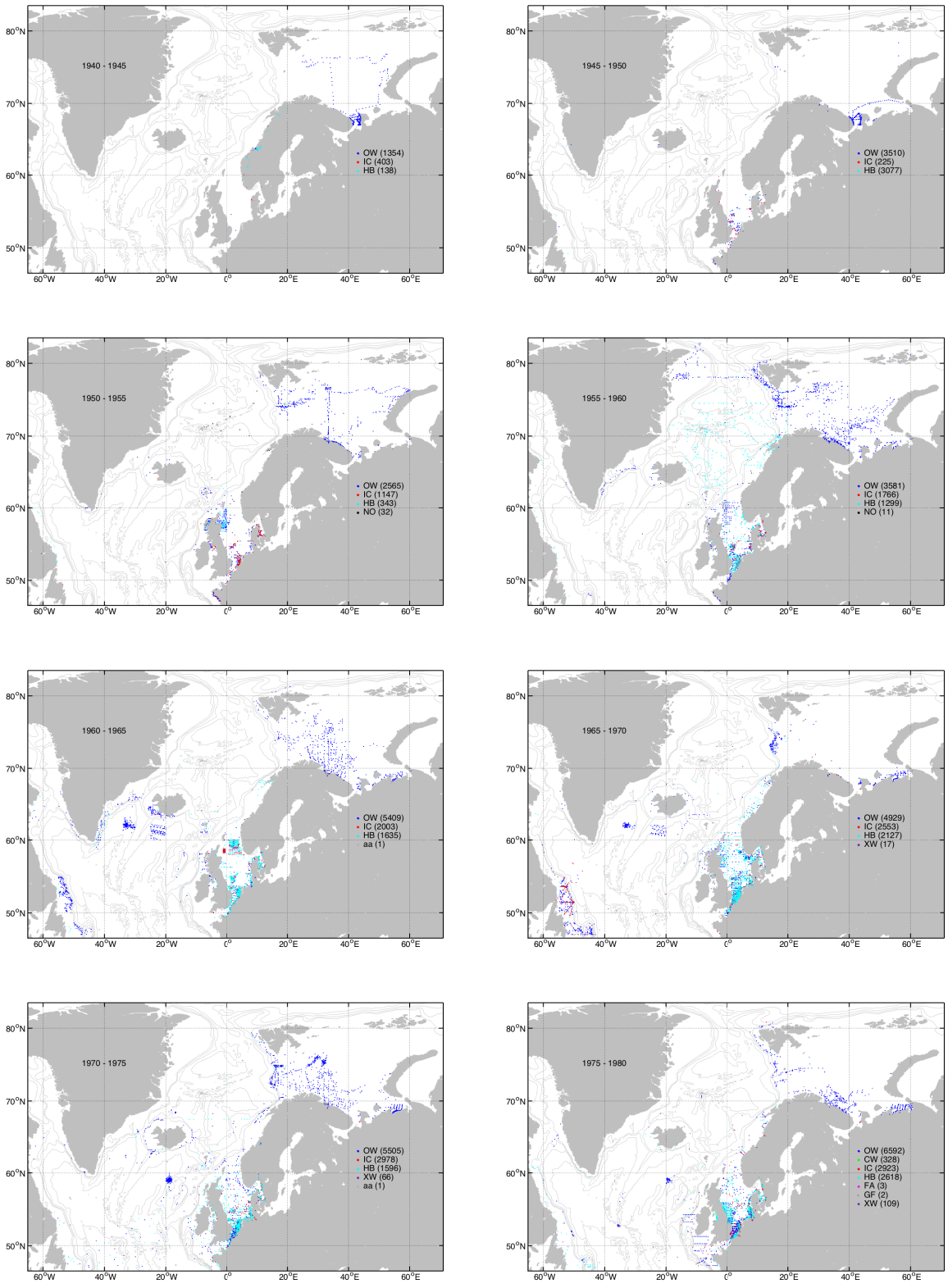


Figure 2.5: Shallow stations in the dataset by pentad (contd.).

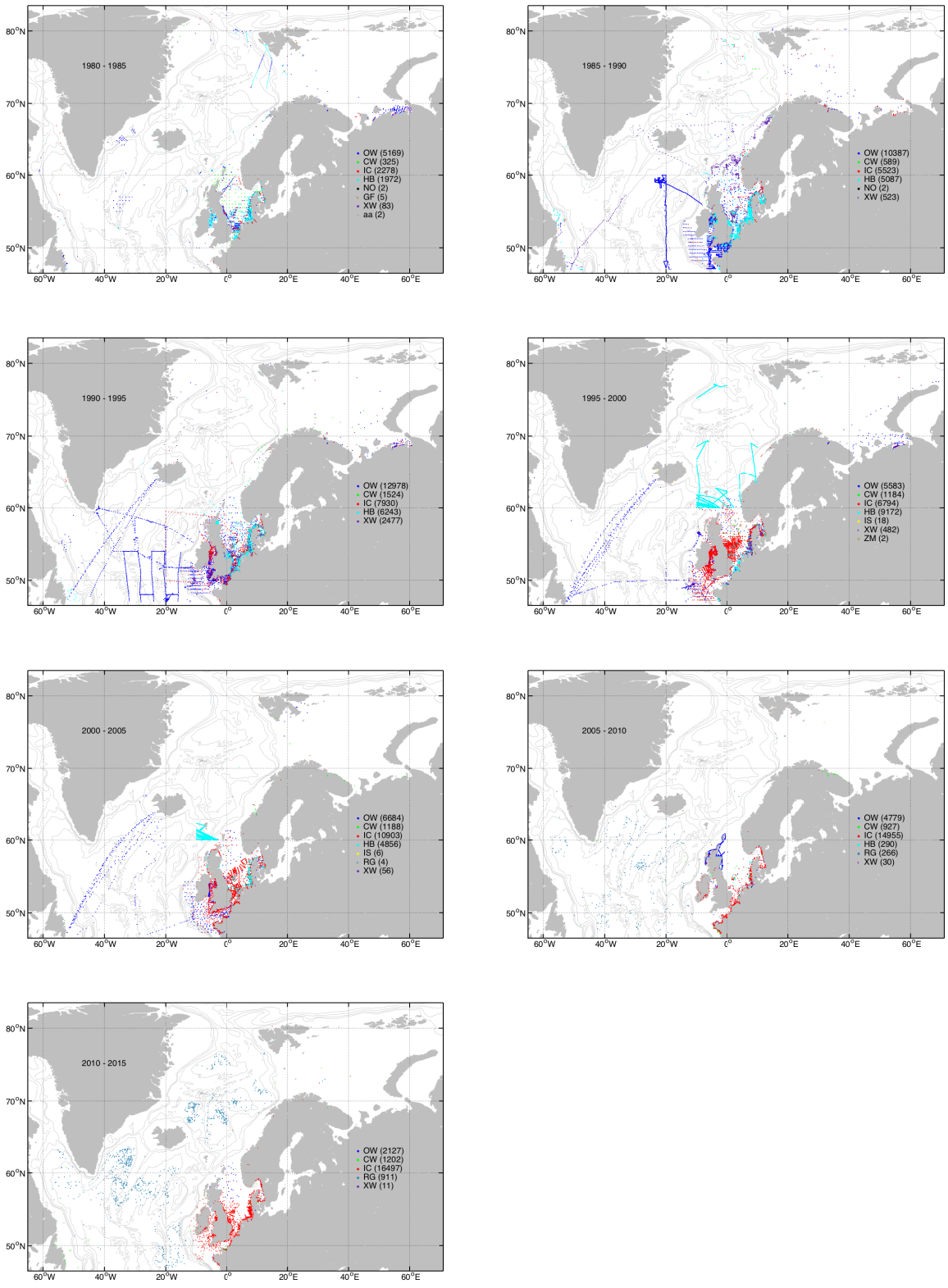


Figure 2.6: Shallow stations in the dataset by pentad (contd).

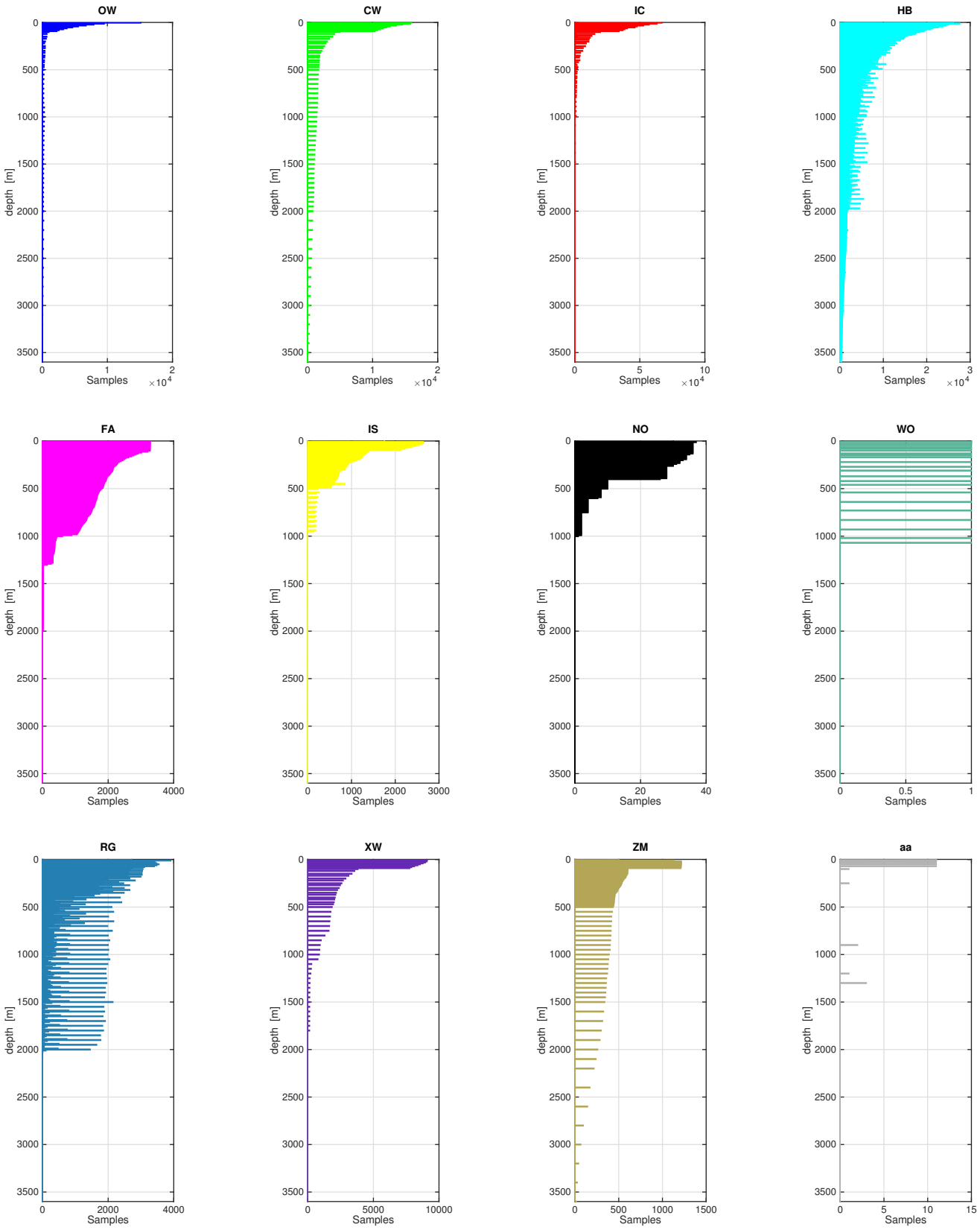


Figure 2.7: Distribution of sample depths from the different source dataset (see Table 2.1 for codes). This example subset is taken from the pentad 2000–2005 (there were no profiles from GF in this pentad).

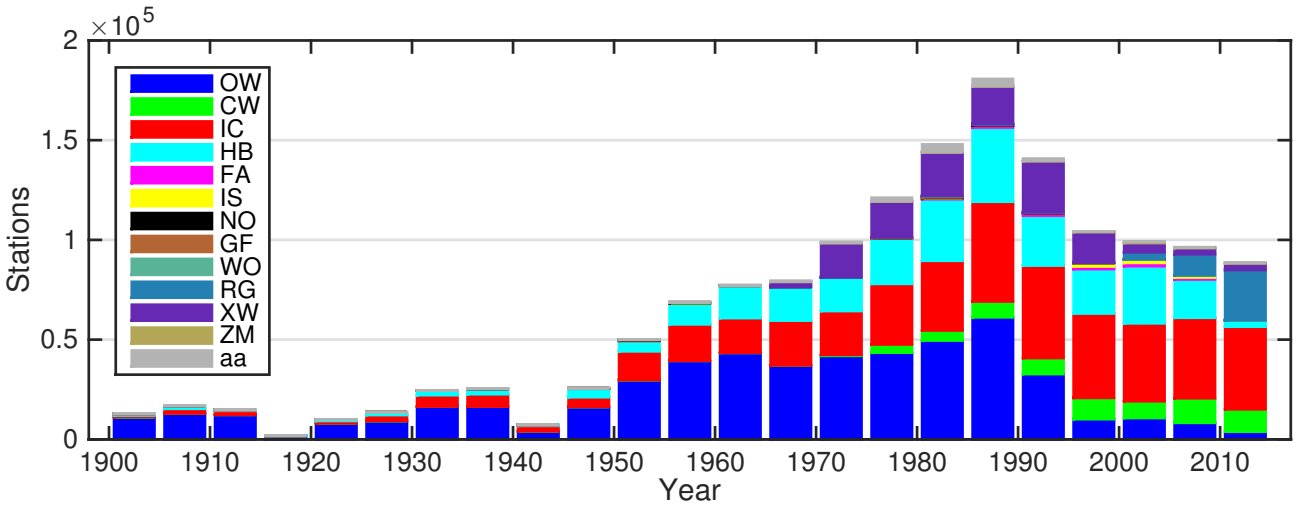


Figure 2.8: Temporal distribution per pentad of total number of stations from each source dataset.

Chapter 3

Data Usage

3.1 Availability

The collection or subsets are available upon request, and all use and publications will have to be reported back, in order to ensure proper acknowledgements are given to the original sources. Contact *even@nersc.no*.

3.2 Terms for use

To fulfill the goals of the projects, it will be necessary to have access to as much as possible of the data acquired through observations. This includes data from several public data bases, but also from sources where access may be restricted. In order to secure the widest possible access without infringing upon originator rights, the following guidelines apply to all data, shared within the project:

- Data exchanged must not be applied for commercial use.
- If a publication relies heavily on a dataset from a specific originator, then the data originator should be included in the author list.
- Any use of data for publication, etc., should bear acknowledgement to the sources for the subset used. See Section 3.3.

The reference to use for technical background is the present report:

> Nilsen, J.E.Ø. (2016). The North Atlantic and Nordic Seas hydrography collection. *NERSC Technical Report* no. 372, Nansen Environmental and Remote Sensing Centre Thormøhlensgate 47, N-5006 Bergen, Norway.

3.3 Acknowledge the sources

When using data from this collection, it is important to acknowledge the originators. Each source database have their own policies and required acknowledgements. In any subset of the NANSHY collection, it is possible to identify the source of all samples and profiles by the `source` variable containing the ID numbers (Table 2.1). Acknowledgements to all sources used must then be given in any publication as follows, with references where available:

1, 2, 11 World Ocean Database 2013 (Boyer et al., 2013). Confer report for details: *data.nodc.noaa.gov/woa/WOD*13*/DOC/wod*13*_intro.pdf*.

3 "ICES Dataset on Ocean Hydrography. The International Council for the Exploration of the Sea, Copenhagen. 2014."

4 Hydrobase3 (Curry and Nobre, 2013). Confer report for details: *www.whoi.edu/science/PO/hydrobase/docs/TechReport*_03*Sep*2013*.pdf*.

5–8, 13 Include by numbers as "Data were provided by: the Faroese Fisheries Laboratory **(5)**; the Marine Research Institute, Iceland **(6)**;, Institute of Marine Research, Norway **(7)**; Geophysical Institute, University of Bergen, Norway **(8)**; the Arctic and Antarctic Research Institute, Russia **(13)**; through the NISE project (Nilsen et al., 2008)."

9 The World Ocean Circulation Experiment (WOCE). Confer website for details: *www.nodc.noaa.gov/woce/*

10 "Argo data were collected and made freely available by the international Argo project and the national programs that contribute to it (Carval et al., 2015)."

12 "Data provided by Centre for Marine and Atmospheric Sciences (ZMAW) during exchange with the NISE project."

Further details on source meta-data can be found in the `Cruise` variable.

Acknowledging the data providers upon publication, is ultimately the responsibility of the user!

Acknowledgments

All data originators mentioned above (Section 3.3) are thanked for making data publicly available. Thanks to Schlitzer (2006) for the ODV software.

Bibliography

Boyer, T., Antonov, J., Baranova, O., Coleman, C., Garcia, H., Grodsky, A., Johnson, D., Locarnini, R., Mishonov, A., O'Brien, T., Paver, C., Reagan, J., Seidov, D., Smolyar, I., and Zweng, M. (2013). World ocean database 2013. In S. Levitus and A. Mishonov, editors, *NOAA Atlas NESDIS 72*. NODC, Silver Spring, MD, pages 1–209. Doi:10.7289/V5NZ85MT.

Carval, T., Keeley, R., Takatsuki, Y., Yoshida, T., Schmid, C., Goldsmith, R., Wong, A., Thresher, A., Tran, A., Loch, S., and Mccreadie, R. (2015). *Argo user's manual V3.2*. Doi:10.13155/29825.

Curry, R. and Nobre, C. (2013). *Hydrobase3*. Technical report, Woods Hole Oceanographic Institution. 37 pp.

Nilsen, J. E. Ø., Hátún, H., Mork, K. A., and Valdimarsson, H. (2008). *The NISE Dataset*. Technical Report 08-01, Faroese Fisheries Laboratory, Box 3051, Tórshavn, Faroe Islands.

Schlitzer, R. (2006). Ocean data view. http://odv.awi.de.