Comment on essd-2022-146
Anonymous Referee #1

Referee comment on "How to learn more about hydrological conditions and phytoplankton dynamics and diversity in the eastern English Channel and the southern bight of the North Sea? the SRN data set (1992–2021)" by Alain Lefebvre and David Devreker, Earth Syst. Sci. Data Discuss., https://doi.org/10.5194/essd-2022-146-RC1, 2022

The authors present a dataset from the shallow coastal waters in the eastern English Channel and southern bight of the North Sea collected from 1992-2001, primarily for the purpose of water quality monitoring, The data are reported to include 281 taxa and 3687 samples spread over 10 stations.

Readers would greatly benefit from a careful and thorough editing of the text, in particular the abstract and introduction. At present ideas are sometimes out of order or disjointed and the purpose of the manuscript is not clear.

The methods and results are quite clear. The overviews of the physical-chemical variables are very helpful. I was surprised to see no indication of dinoflagellates in Fig. 2. Is it possible that there were no phytoplankton dinoflagellates (autotrophs or mixotrophs)? This seems unlikely. I wonder if a seasonal version of the count distribution would be helpful. Are any size data or approximate biomasses of the enumerated taxa available? It might also be interesting to know something of the distribution of species richness, even just the average richness per sample.

I thank the authors for their efforts to disseminate and describe their wonderful dataset. Unfortunately, from my perspective the data are inadequately described. Since this is a data paper in a data journal, this is a problem.

I have examined the data at https://doi.org/10.17882/50832

The data appear to be encoded in latin1 as a csv file with a semi-colon separator. It only
takes a few guesses to work this out, but ideally the reader would not need to guess.

I did not see a data dictionary or other description of the contents of the file anywhere. The variable headings are written in French. I understand the desire to work in one’s preferred language, but the language of publication is English, so a translation should be provided.

I was able to read a data table of 61 variables and 99,006 observations. The number of observations and variables should be reported in the metadata so that the user can be sure the data were received as expected.

There are 6 variables which are missing for all observations. I don’t understand the need to include undescribed variables with no data.

There were some (809) zero counts for taxonomic abundance, but very few (<1%). Please clarify the reason for including these zeros. Was a consistent taxonomic list used for all stations and times? Can the reader infer that the taxa recorded at some stations but not others have zero abundance when not reported?

It would be helpful to provide latitude and longitude of the stations; these can be read approximately from Fig. 1, but they do not appear to be in the data file. I was unable to decode the station location data: cordonnees passage min and max for x and y.

The values I computed for Table 2 did not match the authors’, likely because I misinterpreted something; incomplete description of the data makes this easy to do unfortunately. I suggest additional details to clarify the differences.

Is there a problem with station SRN Somme mer 1 (Mer1?) resulting in it not being reported in Table 2?

Number of observations (samples) per station:

```
z1 |> count(lieu_de_surveillance_libelle, passage_identifiant_interne, passage_date) |> count(lieu_de_surveillance_libelle)
```

# A tibble: 11 × 2
lieu_de_surveillance_libelle n

1 At so  479

2 Bif  414

3 Mimer  272

4 Point 1 Boulogne 508

5 Point 1 Dunkerque  406

6 Point 2 SRN Boulogne 401

7 Point 3 SRN Boulogne 395

8 Point 3 SRN Dunkerque  324

9 Point 4 SRN Dunkerque  309

10 SRN Somme mer 1  301

11 SRN Somme mer 2  391
The total number of samples in the dataset is 4200. This does not match Table 2 (3687) even if all the observations from SRN Somme mer 1 are removed. My calculations showed that 2007 had the most observations (179), but this does not agree with the number in the abstract (184, line 12). It would be more representative to report the mean or median number of observations (142, 140) or the range (100 to 179).

Number of species:

```r
z1 |> count(resultat_nom_du_taxon_referent, lieu_de_surveillance_libelle) |>
    count(lieu_de_surveillance_libelle)
```

# A tibble: 11 × 2

<table>
<thead>
<tr>
<th>lieu_de_surveillance_libelle</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>At so</td>
<td>224</td>
</tr>
<tr>
<td>Bif</td>
<td>206</td>
</tr>
<tr>
<td>Mimer</td>
<td>197</td>
</tr>
<tr>
<td>Point 1 Boulogne</td>
<td>215</td>
</tr>
<tr>
<td>Point 1 Dunkerque</td>
<td>221</td>
</tr>
<tr>
<td>Point 2 SRN Boulogne</td>
<td>208</td>
</tr>
<tr>
<td>Point 3 SRN Boulogne</td>
<td>188</td>
</tr>
</tbody>
</table>
8 Point 3 SRN Dunkerque 202

9 Point 4 SRN Dunkerque 191

10 SRN Somme mer 1 167

11 SRN Somme mer 2 197

Has the taxonomy be standardized in any way? Was a database such as marinespecies.org used? I suspect the count of species in table 2 in fact refers to some level of taxonomic resolution and not species. In addition to species, the data table reports many genera, some higher classifications, and size or shape features. Some taxa are fusions of several species or genera, e.g., “Chaetoceros densus + eibenii + borealis + castracanei”. Some classifications can be guessed, but are incomplete, e.g., “Centriques”, “Pennées”.

I did not see any spelling errors in the taxonomic identifications; I congratulate the data curators for this success!

Phaeocystis globosa was repeatedly identified in the manuscript, but does not appear in the dataset. Only the genus-level id Phaeocystis is reported in the data. This is a serious oversight or inconsistency. It would be helpful to note in the data if the counts refer to cells, colonies, or a mixture.

I was able to read the physical-chemical data and interpret it. The general concerns above copy over here about documentation, encoding, location information. I did not see any information about the units of measurement of the various quantities (temperature, chl-a, nitrate, nitrite, phosphate, silicate; some can be guessed, but the nutrients could easily be in mass or mol units and there is no way to tell.) Guessing is not ideal in a documented dataset. The metadata indicated phaeopigments, suspended matter organic and mineral are reported, but I did not see any observations of these quantities in the data. These are oversights that should be corrected.

I did not examine the data at the following sites as they did not seem to be the primary target of this data paper:

A bit more information at line 347 about the relationship between these data would be helpful. (Are they completely distinct, partially overlapping, etc.?)

I was unable to use the R package TTAinterfaceTrendAnalysis. It requires X windows and Tk application software which many users will not have installed. It might be helpful to indicate something of the software requirements in a brief note. These requirements are a bit unusual for modern software packages and will likely limit the usage of their package.

Detailed comments

Abstract

The abstract does not clearly describe the dataset, which is the main purpose of this paper. I suggest informing the reader of the years covered by the data and the total number of observations.

The abstract is written about an “historical” dataset, yet is largely written in the present tense (SRN collects, objectives … are, regular acquisition of data…) This is a bit confusing, indeed the paper both describes an ongoing program and presents data from 30 years of collection. A bit of smoothing of the exposition and clarification of the goals of the manuscript would help the reader.

Line 8. Define acronym SRN before it is used (It is defined repeatedly below)

Line 8. Give location of Iframer (Brest)

Line 8. What makes the data historical? Has data collection ceased? Are modern data excluded?

Line 12. 184 samples in one year would be quite intense sampling. Is this phrasing correct? Samples per station and the number of stations might be more informative.

Line 14. Are continental inputs, development and management policy metadata part of the time series? Are these data described in this manuscript? Available publicly?
Introduction

Line 30: “others cause excessive organic matter inputs”. I think I know what you mean here, but this is a relatively unusual observation, so an example taxon or citation could help make your point more clearly.

Line 59. What does “address” mean here? It’s a fairly indirect verb.

Line 74. Presumably French should be capitalized here.

Purpose

Line 97. Should “propose” be “present”? 