

# Response

H. Sherry Zhang, Dianne Cook, Ursula Laa, Nicolas Langrené, Patricia Menéndez

2023-03-04

## Response to B-review\_\_1\_\_1

### The Paper

The package and the proposed data structures could certainly constitute a valuable base for the spatio-temporal statistics tooling. Unfortunately the paper does not clearly describe the data structure, nor the associated functionality. I had to read the package vignettes, run the attached code and even look into the package's code in order to fully understand the data structure.

#### Re:

Given the small size of the package, the paper feels a bit on the heavy side. Most of the code in the paper is not self-contained which makes it rather difficult to follow. Examples at the end of the paper don't elucidate the inner workings of the package but rather advertise the functionality of other, mostly visualization, packages.

#### Re:

More fundamentally, the design of the package has an immediately apparent drawback. The `cubble` class is used to represent two incompatible data types - long and nested. Internally, the two data types are distinguished by the internal attribute "form". In other words, the authors use an attribute of an object to mimic the functionality of a sub-class.

#### Re:

Relatedly, the rationale for the parallel naming convention long/nested and temporal/spatial is not entirely clear. It seems to me that the semantically unambiguous temporal/spatial could be used throughout without ambiguity, thus resolving the terminological redundancy.

#### Re:

My recommendation would be a combination of the following:

- Rework the definitional parts (sections 2 and 3) by following more closely the "design" vignette. **Re:**
- Describe at a glance the core functions which operate on the class and their motivation. **Re:**
- Rework section 4 by making it more concise and provide links to sections in supplementary material or dedicated vignettes. **Re:**
- Rework examples section 5 into "Applications". Briefly describe the applications and provide and refer to dedicated and self-contained vignettes. **Re:**

### The Package

The package is generally badly documented. The meaning of the arguments is almost never clear from the documentation alone. For example the documentation of `as_cubble` states:

- What objects are supported as input data?
- What is the assumption of the "nestedness" of the input data.
- What is the accepted data type of key, index and coords?

**Re:**

## Further suggestion

Consider not throwing when `face_spatial` is applied to the spatial object. This would allow for generic code where the “form” of the input the data is not known or does not matter.

**Re:**

It would be useful to print a few rows of the spatial attribute of the “long” object.

**Re:**

Section 3.6 is missing `tidyr` package which is especially relevant here given its nesting and unnesting operations <https://tidyr.tidyverse.org/articles/nest.html>

**Re:**

More comments inlined in the pdf.

- on p4 “The following sections explain their roles, why the new cubble class is needed and how the package relates to existing packages for spatial and temporal data analysis.” The paper is a bit on the verbose side. Such comments could be safely removed. **Re:** sentence removed.
- on p5 section 3.1, “The arguments `key` and `index` follow the wording in the `tsibble` package to describe the temporal order and multiple series while `coords` specifies the spatial location of each site.”, Unclear what **key** and **index** mean. Isn’t `key` a spatial identifier and `index` a temporal one? **Re:**
- on p5 section 3.1, the word “wording” is highlighted in the sentence “The arguments `key` and `index` follow the wording in the `tsibble` package to ...”, conventions. **Re:** *wording* changed to *conventions*
- on p6 section 3.2, ‘The cubble class also provides a long form, which expands the `ts` column and temporarily “hides” the spatial variables.’, confusing: **hides** means storing it as an attribute. **Re:**
- on p6 section 3.2, I think it would be a good idea to clearly describe the two data structures in an earlier section. By this point it’s still not clear what those are and what the meaning of `key` and `index` is. **Re:**
- on p6, section 3.2, yet another reason to call this data structure “`cubble_temporal`” instead of the semantically unobvious “long”. **Re:**
- on p8 section 3.5, Isn’t the root reason the memory efficiency? That is, the spatial component cannot be represented as part of the nested tibble without repetition. **Re:**
- on p8 section 3.6, The missing elephant is `tidyr` which has nesting and unnesting operations <https://tidyr.tidyverse.org/articles/nest.html>. **Re:**

## Response to B-review\_\_2\_\_1

### The Manuscript

The manuscript is in general clearly structured. However, an important reference to the R-package `sftime` is missing that also deals with the representation of spatiotemporal data. It might be beneficial to provide illustrative examples in the manuscript that clearly compare between a cubble representation and representations of existing packages. Which data sets/structures cannot (or with a greater effort) be represented with existing R-packages such as e.g. `stars` and `sftime`?

**Re:**

For spatial data, the coordinate reference system is essential metadata information. In one example of the manuscript, the input data does have information on the CRS but it is not discussed in the manuscript how that is handled within cubble. How would coordinate transformations be handled with cubble?

**Re:**

The use-case of temporal matching based on features of the time series appears an interesting, but also a very specific one. How could this be generalized to a more generic approach?

**Re:**

Additionally, several typos and language issues arise and limit the readability of the manuscript. Some examples are:

- doubled/missing words/none correct sentences:
  - “*components* spatio-temporal *components*”
  - “... fits works ...”
  - “... be activate rows ...”
  - “... highlighted *in* the ...”
  - “. . . using 2020 measurements using `match_sites()` function.”
  - “An example of this using is included in the Appendix”
  - “. . . the data in -a- multiple -of- ways on-the-fly”
  - **Re:** the mentioned examples have been changed to: “... on different spatio-temporal components”, “The workflow ... works well with ...”, “... will activate rows having ...”, “... highlighted in the map”, “... using the `match_sites()` function.”, “An example of this is included in the Appendix”, and “... the data in multiple ways on-the-fly”.
- Surprising references: in Section 4.4 it says “. . . Glyph maps (Section 3.4)”, but Section 3.4 says “. . . glyph maps will be explained in Section 4.4”. **Re:** the reference to Section 3.4 in Section 4.4 has been replaced with the literature (Wickham et al. 2012).
- Should “. . . it’s temperatures are more *consistent*” rather be “. . . it’s temperatures are more *constant*”? **Re:** changed to “more constant”
- Typos:
  - “. . . spatial and *tmeporal* information are available.”
  - The polar vortex, signalled by the high *speicfic* humidity, splits into two on 2002-09-26 and *further-s- split\_s* into four on 2002-10-04.
  - the data in -a- *multiple -of-* ways on-the-fly.
  - **Re:** typo corrected to “temporal”, “specific”, “further splits”, and “multiple ways”

## The package

I have been a bit puzzled by the “print” of a nested cubble in its temporal face. The row # temporal: date [date], prcp [dbl], tmax [dbl], tmin [dbl] appears at first sight (also in comparison with the print of the spatial face), as if the temporal domain is given by all those variables. To me, a notion such as

```
# temporal: date [date]
```

```
# variables: prcp [dbl], tmax [dbl], tmin [dbl]
```

would have been more intuitive. Possibly also including the temporal range (as for the spatial domain its bbox).

**Re:**

In the manuscript and in the reproducible-script.R, a data set `historical_tmax` is introduced. In the package, the corresponding data set seems to be `tmax_hist`.

**Re:**

The manuscript cannot be reproduced, as the reproducible-script.R (part of the submission to JSS) only uses a subset, but does not provide a separate result file that would allow to compare the reproduced outputs with the desired output for the subset.

**Re:**