Geoscientific
Model Development
Discussions

# Interactive comment on "Silicone v1.0.0: an open-source Python package for inferring missing emissions data for climate change research" *by* Robin D. Lamboll et al.

**Anonymous Referee #2**

Received and published: 2 July 2020

The paper by Lamboll et al. describes an open-source python package, Silicone, which comprises a collection of algorithms to derive emission pathways of gasses missing from a certain data source using the pathway of an available gas (lead gas) together with the pathway of the available and the missing gas from other data sources. The core of the paper is the description of the currently implemented algorithms (termed crunchers) together with a guide for when to apply which of these crunchers. In addition, the authors try to derive a recommendation for which gas to use as a default lead gas and they demonstrate the applicability of the tool on different test/use cases.

Overall, I find the paper in large parts fluent and well-structured and, albeit I find this collection of algorithms a rather small scientific contribution in itself, it seems to be a self-contained part of a larger pipeline for climate assessments and therefore probably a valuable contribution for climate research within the scope of GMD.

**General comments**

1. In most parts of the manuscript and in the source code the specified aim is to complete missing emissions in future scenarios produced by IAMs.

a) However, in some sections of the manuscript, other aims are indicated (e.g. complete stylised scenarios, fill missing sectoral data, use historical estimates, aggregate regional data(?)) which I found confusing sometimes. One solution could be to remove these hints of other applications and use the (IAM/stylized) future scenarios as the aim and example throughout the paper. Additional applications could then be outlined in a discussions section (in a bit more detail).

b) With respect to the main aim I would have expected use cases showing the completion of several variables of different IAM scenarios.

2. I would recommend to rework parts of the abstract and of the introduction in order to better cover the content/ set the scene for the rest of the manuscript (e.g. add info about rank correlation, tests/ use cases).

3. Literature: it is rather difficult for me to imagine that there are no other somewhat comparable tools around and that so far missing emissions were usually set to zero or only somehow unsystematically filled following 'ad hoc' decisions (as stated in the introduction). For completeness it would be nice if the authors could dig some more into the literature and check how climate models so far got the required input from IAMs? One example for a tool covering a similar purpose in maybe a slightly different but connected setting is the tool used in Gütschow et al. (2016) and du Pont et al.

(2016) which is described in Nabel et al. (2011). Some of the co-authors have been involved in these papers.

4. While reading I sometimes got confused by different terms and I found parts of the manuscript a bit sketchy or difficult to read. More specific:

a) There are several changes in terminology among different (sub-) sections (e.g. "lead variable", "lead gas" "inputs or outputs" (l.89), and timeseries; and a sudden switch to model and scenario in 2.2.2 where 2.2.1 only had the more general term database; but also small things as the change from CH4 in section 1 and 2 to methane starting section 3). I think for the reader it would be helpful to stick to a certain terminology throughout the manuscript.

b) The terms infiller and infilly are very difficult to distinguish in quick reading and I think it would help a lot when choosing less similar terms – How about source and sink/target, or infiller and target, or infiller and silicon-filled, or comparable.

c) In subsection 2.2.2, a bit out of the sudden, several scenarios and models seem to be presupposed, while at the beginning of section 2 only "a database that contains data for at least two emission species" is kind of officially introduced. Maybe it would help to directly introduce the use of different models/IAMs and their scenarios at the beginning of section 2 such that the usage of different timeseries (2.2) and different models and scenarios (2.2.2) is less out of the sudden. An alternative could be a consistent use of the more general terms of "timeseries of different lead variables in the infiller database" depending on the main aim of the paper (see also point 1).

d) Please consider to better structure 2.2.1 and 2.2.2 (e.g. print algorithm names in bold or in italics, with separated paragraphs for the different algorithms or /and as lists (e.g. latex 'description' or the like)).

e) Equations are throughout embedded in the text (see also point 5).

5. Since the different algorithms for the completion of emission timeseries are the

main scientific contribution of the paper I would appreciate if the equations could be clearly separated from the text (i.e. introduced as separated numbered equations) and, furthermore, if more equations would be added (see also specific comments below). In my opinion this could increase readability (4e) and reproducibility. Ideally it could also help to better understand how cases of several lead variable pathways are treated in the different algorithms.

6. Test and use cases only show the usage of absolute value based algorithms, which I find unfortunate and a bit incomplete given the share of the method section dedicated to the ratio based algorithms. How about at least including examples using the time dependent ratio method?

7. Please check the format of your references in the text (e.g. l.28, l.59, l.60, l.68, . . . )

**Specific comments/questions**

l.1 Why Silicone?

l.16 Transition. E.g. "In this paper. . ."

l.16 Please consider to add more information here about the content of the paper

l.33 . . . exert . . . between? Please check language

l.36 Is there an example reference/study where filling with zeros has been applied?

l.41 I do not understand "does not scale easily"

l.45 There is no 3.8.1 in this reference (reference currently points to Chapter 4, if you target 3.8.1 consider citing Teske et al.)

l.60 Please specify what "this" refers to

l.64 "suite of tools" are these all python tools?

l.66 Are there more than these two? Else consider to add "and/as well as" before

"harmonise..."

l.67 Consider deleting ", managed by the OpenSCM community"

l.82 several? Three/two...

l.85 Maybe change to "Currently, there are ..."?

l.98 Consider rephrasing e.g. "... and where emissions are expected to scale with each other ..."

l.98 What do you mean with "regional data" and "aggregate data"? If you refer to regions as subset of global data then this is the first time that a spatial reference is given and I wonder if it would be appropriate to introduce this more formally earlier in the manuscript?

l.100 Consider to give an example? CO2 uptake?

l.106 What do you mean with similar – similar magnitudes?

l.111 "estimate the ratios" – if not to be predefined...

l.112 "follower value in infillee database" -> "in the ..."

l.112 Please consider to visually separate (and number) the equations.

l.114 "each different timeseries" -> different regarding what - do you mean different follower variables?

l.116 mean regarding what - time or different sources (models, scenarios?)

l.117 what does "both" refer to?

l.119 what do you mean with "all estimates"? Different sources (models, scenarios)?

l.120 why historical? Couldn't this also be different scenarios from different IAMs?

l.120 "and the lower case $e_f(t)$ represents the follower values in the database at time

t." -> "and the lower case $e_f(t_{last})$ represents the follower values in the database."?

l.125 the infillee lead is not part of the formula – is this the $E_l(t)$ from l.113?

l. 125-133: Maybe consider to restructure? You could start with the context, i.e. the algorithm name, and then add the explanation, e.g.: "The decompose collection multiple infiller is based on... relying on the useful property of ..."

l.134-137: Equation for R(t)?

l.142 I did not understand this sentence

l.143-145: Equation for $E_f$?

l.151 There is no 3.8.1 in that reference (reference currently points to Chapter 4, if you target 3.8.1 consider citing Teske et al.)

l.152 There seems to be a lost copy of the figure caption in the text. (Either just delete or maybe rephrase to steps with complete sentences)

l.155-165 Equation for $E_f$?

l.163 Maybe $E_f$ not $E_l$?

l.166 What about the KyotoGHGs as one basket?

l.169 Maybe give an example for two such variables?

l.185 Consider explicitly listing the two constituents

l.186 Which are "these two" – CO2 and CH4?!

l.188 you write that BC, CO and OC "correlate poorly with others, however, from the table it seems that they do not correlate less well with others than other gasses, the main difference is that they correlate very well with each other.

l.190 maybe aggregate F-gas emissions / F-gasses as a basket?

l.191 up to here always "CH4"

l.198 consider deleting "and find similar results"

l.199 "we choose four" – which are basically all?

l.199 But even if there are errors, wouldn't it be interesting to see what happens? I would appreciate if you could also show results for the 'time dependent ratio' algorithm

l.201-202 Equation?

l.203 "both cases" – CO2 and CH4?

l.203 "non-CO2 pathways" – but CO2 is derived with CH4 -> maybe replace by "emission pathways"

l.204 I would rephrase this, if QRW would be fairly similar all four would be?

l.208 What do you conclude from the non-Gaussian distribution test?

l.212 "either of CO2 or CH4"

l.214 capital T for Table 3

l.225 Add Silicon -> "Data in the Silicon package"

l.238 treatment of regions has not been introduced, maybe explain better or consider deletion?

l.243 "this database" – which? The SR1.5 repository?

l.249-252: Again a lost copy of a figure caption

l.285-287 "free variables" are mentioned twice but are not further explained?

l.287-288 I do not understand this sentence

l.300 consider deleting "of which there are many" or maybe replace by "several options" or the like

Table2: Please explain the asterisk again in this figure caption

Table3: Consider to change the colouring – to me the yellow/orange highlighting gives a 'positive' impression. Maybe you could colour the cells with bold numbers in green and those which are currently yellow in red?

**References**

du Pont, Y.R., Jeffery, M. L., Gütschow, J., Christoff, P. Meinshausen, M. National contributions for decarbonizing the world economy in line with the G7 agreement. Environ. Res. Lett. 11, 054005, 2016.

Gütschow, J., Jeffery, M. L., Gieseke, R., Gebel, R., Stevens, D., Krapp, M., and Rocha, M.: The PRIMAP-hist national 1208 historical emissions time series, Earth Syst. Sci. Data Discuss., 2016, 1-44, 10.5194/essd-2016-12, 2016.

Nabel, J. E. M. S., Rogelj, J., Chen, C. M., Markmann, K., Gutzmann, D. J., Meinshausen, M.: Decision support for international climate policy–The PRIMAP emission module. Environmental Modelling Software, 26(12), 1419-1433, 2011.

Interactive comment on Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2020-138, 2020.