Review of “jsmetrics v0.1.1: a Python package for metrics and algorithms used to identify or characterise atmospheric jet-streams”, by Keel, et al.
Reviewed by Gloria Manney

Recommendation: I raise some questions about the maturity of the jsmetrics package/documentation and thus the current suitability for publication in GMD. Substantial revisions and clarifications in the manuscript, and improvements in the code documentation, would be needed before publication.

General Comments: I was initially very excited to read this manuscript and try the jsmetrics package. I think the idea of putting together this sort of package is a great one. After reading the paper, installing (via pip, so presumably latest stable version as of late July 2023) and trying out the package, and looking at the documentation both on GitHub and the site given in Keel (2023), I have several reservations about this package as it now stands regarding its maturity, its interpretation and implementation of some of the methods, the frugality of the information returned by the package, and the opaqueness of the documentation and usage guidance. In particular,

(1) Of the paper types listed as suitable for GMD, the category this would fit into is “methods for assessment of models”. The description of this paper type includes “..description of a fully fledged software tool…”. I question whether this package can at present be considered “fully fledged”. For example, of the 17 metrics (summing the three types) described and exemplified in the paper, on the website, four of them are listed as complete, one as “in progress” with a note indicating “need help”, and the rest as “to verify”. As far as I can tell (e.g., extrapolating from Fig. 1 in the paper), “to verify” means that the implementation has not yet been verified against results from the original paper. Even if thus limited only to the approaches discussed in the paper (where the reader is likely to assume that they have all been fully tested), I question whether this can be considered a mature software package. I recognize that the authors are hoping to make this a community project and entrain outside contributors, but if this manuscript is a call for participants – without a basic foundation of mature software that is to be built upon – then I am not sure it is appropriate for publication in GMD at the present state of maturity of the software package.

(2) I question the distinction (and the need for one) between “jet statistics” and “jet core algorithms”. All of these methods aim to extract basic information such as the jet latitude (and altitude for several of them) and windspeed. For example, the method of Manney et al. (2011, hereinafter M11; though the package is formally named JETPAC, that was not mentioned in the first couple of papers) was based on extending that of Koch et al. to characterise the jets in the latitude/altitude plane rather than only getting horizontal locations and vertically averaged dynamical characteristics. But the two papers (as well as Manney et al., 2014) have very similar aims and show very similar diagnostics (eg., frequency distributions of jet cores and associated winds) – so why is one “jet statistics” algorithm and the other a “jet core algorithm”? I see no need for a distinction here.

(3) Obviously, there is one method herein with which I'm very familiar (and to a lesser degree several of the other methods, either because mine built on them, they (apparently) built on mine, or I reviewed the original papers. I looked closely at their implementation of the M11 method, and I cannot from the information provided (nor from running an example of their implementation) convince myself whether or not what they are doing closely resembles the M11 method: In the paper (lines 137–140)l they do not list M11 in the category "(ii) in relation to wind-speeds of neighbouring data points (local wind-speed maxima…", but this method is fundamentally based on first finding all local wind-speed maxima (in a latitude / altitude plane) and then applying thresholding and other criteria. In fact, the M11 method is fundamentally nearly identical to the Peña-Ortiz et al (2013; hereinafter PO2013) method (the latter paper showed internal evidence that they must have known of M11, but did not cite it), which is listed in both categories that would be appropriate for M11.

In what appears to be the primary function to identify the cores, there is a comment saying "Core are discovered where 8-cells are above boundary threshold… Paper uses 100-400 hPa." The M11 method inherently operates on a slice of the windspeed (and other fields) with a latitude-like and an altitude-like coordinate (and then is generally run at each longitude gridpoint for each day); jet cores are identified in that latitude/altitude plane by finding the local maxima (as points with higher windspeed than the nearest neighbors in altitude and latitude) and testing for exceedance of the core threshold. Finding the "edges" of the jet regions is no more complicated (or informative!) than finding the latitude gridpoint on each side and vertical coordinate (eg, altitude) gridpoint above and below where the windspeed drops below the boundary threshold value. Other criteria that determine whether to consider two cores in the same "region" (that is, where the windspeed does not drop below the boundary threshold between them) separate are latitude distance between the cores (which this package appears to implement) and the amount by which the windspeed drops along a line connecting the cores in altitude / latitude space (which I didn't find mentioned herein or in the documentation). 100-400 hPa is the region to which the jet **core** altitude is limited, but the jet regions surrounding any core are identified at altitudes that may be above and/or below that – wherever they occur. This procedure is described in detail in M11 and Manney et al (2014), and completely but in less detail in Manney et al (2017, 2021; latter J Clim, DOI: 10.1175/JCLI-D-20-0947.1) and Manney & Hegglin (2018). It is not clear to me that either the description in this manuscript and the jsmetrics code, or the implementation in the code are doing something that is fundamentally similar to this method.

Further, in lines 155–157, they ignore the update to the M11 method provided by Manney & Hegglin (2018) (and used in all succeeding papers using the method, including Manney et al, 2021), whereby the subtropical jet is identified (and thus distinguished from polar jets) using a physically-based method. In the preceding lines, they also fail to note that PO13 themselves concluded that their method for separating subtropical and polar jets did not work well; which is consistent with the statement in Manney et al, 2014 regarding the division of subtropical and polar jets by latitude (which they used in the one case where they did not simply analyze all the jet cores at each longitude) being useful only for very broad climatological studies.

I've gone into (probably tiresomely) lengthy detail here because this case (and the difference in the descriptions of M11 and PO13, which are very similar methods) makes me question whether there may be other metrics for which the implementation may not accurately reflect the original method. If I had been putting together a package of this sort, I would have contacted the authors of the original papers and tried to verify my understanding of the metrics and what was fundamental to them – has this been done for any of the methods? You probably would not have gotten a 100% response rate, but many authors (like me) would have been happy to confirm the algorithm and to support your efforts to the extent we could find the time.

(4) The information that is provided as the output from running these algorithms does not include some of the most fundamental outputs. Continuing to use M11 as an example, the fundamental output provided by M11 (ignoring their very complete characterization of numerous other dynamical fields at the jet core locations) is not just jet core latitude, but also jet core winds (windspeed, zonal and meridional winds), and jet core altitude (PO13, at least, also provide these characteristics, and I presume others do as well); in M11 the characterization of the boundaries of the jet regions is secondary information, they focus on full characterization of the jet cores. These are all very fundamental quantities that are used in virtually all studies of jets and their trends (omitting altitude for methods that don't provide that – but for methods that do it is one of a several fundamental characteristics that are expected to change with climate change, see, e.g., discussion / references in Manney & Hegglin 2018) and in most cases (certainly for M11) these quantities "come along" with (because they are a fundamental part of) the jet core identification / characterisation. Thus it seems to me an important (and easily remedied) omission to not provide these quantities in the output.

(5) I find the documentation – in the GitHub project, in https://jsmetrics.readthedocs.io/en/latest/, and in the code itself – opaque, hard to navigate, and incomplete (though the last may be an artifact of the previous issues). Indeed, part of point (3) above could be because I could not find complete descriptions of the algorithm as implemented for each of the methods. Another example is the "Usage" information, which says that to run the Kuang et al (2014) algorithm, after reading/formatting a u & v wind dataset in uv_data, you would type:
k14 = jsmetrics.jetstream_algorithms.kuang_et_al_2014(uv_data)
Whereas I had to type (syntax obtained from looking through the code):
k14 = jsmetrics.metrics.jet_core_algorithms.kuang_et_al_2014(uv_data)
(I cannot find any objects with names containing "jetstream_"). I did, finally, find some end-to-end usage examples (that have correct syntax) by following several layers of links from the main README.rst file in the GitHub project – this information should be in https://jsmetrics.readthedocs.io/en/latest/ (in a corrected / expanded usage section, and either in an examples section like I've seen in other packages I've installed with this format of documentation, or with the API reference information as examples for particular metrics / algorithms).
In "Metrics & Algorithms" in the documentation (link above):

- The "all metrics" link gives a 404 error.

- There do not appear to be any API-type descriptions of the “Metrics”, thus zero guidance on how to run them; the API-type descriptions of the “Algorithms” are not complete enough to be useful. For instance, basic information for formatting the input fields (whether the algorithm needs one level or multiple levels; if it works on one level at a time, does it handle, and how does it handle, multiple levels) should be given. There is also little or no information about what the output Dataset contains in each case. There should be a full API reference that describes each routine in enough detail that a user with basic familiarity with python can run them. (I ran several of them, but generally had to go to source code and even then guess at some things to see how to do that.)
- In the API-type information given for the “Algorithms”, the descriptions of the methods are not clear (see discussion re M11 above), nor sufficiently detailed to know what the implementation actually does. The code itself, for the vast majority of routines that actually do things (as opposed to just set up inputs, etc) are going to be very opaque to anyone who isn’t fluent in the full object-oriented syntax / usage – that class includes many, many python users in the atmospheric science / climate community (partly because python can be used so effectively without even getting to using it in true object-oriented implementations). I’m not criticizing how the code is written, but doing it this way makes it critical to document what it actually does so the user doesn’t have to be able to read the code at that level. (The page at the “issues” link gives descriptions of the algorithms directly from the original papers, but nothing that shows how that is “translated” into the implementation here. Being able to see the pseudo-code that is mentioned in the paper would be extremely helpful as one way to do this.)

Further details are given in the specific comments. Altogether, the package and its documentation appear sufficiently immature to question “announcing” it in a paper.

Specific Comments (in order of appearance in the manuscript, not importance):

Lines 1–17, The abstract would benefit from spending less time on motivation and more on describing the actual results of the paper (just 3 / 17 lines are devoted to what the paper actually does).

Lines 26–33, See general point (2).

LInes 53–55, Should cite Lee & Kim (2003) and Manney et al (2014) in addition to Spensberger and Spengler (2020) (indeed, much of the latter’s discussion was based on the discussion in those earlier papers).

Line 81, not all of these (eg, Barton & Ellis use one specific pressure level) describe the “atmospheric column”.

Lines 91–92, these also do not capture the behavior near the level of maximum wind speed, see, e.g., discussion in Melamed-Turkish et al. (2018) and Manney et al (2021).

Lines 93–95, how is “extracting the latitude of the jet stream as the point of fastest zonal wind” different from setting “a value of jet latitude and jet speed for each longitude”? Are you saying in the first statement that the maximum is found as a function of latitude and longitude? (Note also, per comment on metrics versus algorithms, if you add “and jet altitude” the latter accurately describes the M11 and PO13 methods, and probably others, that are categorized as “jet core algorithms”.)

Line 125, since this is indeed a “highly contested” topic, it might be good to cite some papers led by other authors, e.g., Manney & Hegglin 2018 have a fairly detailed paragraph in their introduction, citing numerous still-relevant references; Francis (2017, BAMS, DOI:10.1175/BAMS-D-17-0006.1) would also be appropriate.

Table 3 and discussion thereof, it would be helpful to distinguish between methods that provide jet altitude as well as jet latitude information (eg, M11, PO13) and those that provide only jet latitude information and that often over a vertically-averaged range (eg, Koch et al, Archer & Caldeira, Schiemann et al, Kuang et al)

Lines 139–140, per general comment (3), M11 also fall into category (ii)

LIne 142–146, It could be just the wording, but this reads as if you are saying the methods of Manney et al (including the important refinement in Manney & Hegglin, 2018 of a physically-based method to distinguish subtropical and polar jets) does not “resolve” different reasons and seasons, when in fact they specifically focused (particularly Manney et al, 2014, 2017, 2021 and Manney & Hegglin, 2018) on analysing regional and seasonal variations. Manney & Hegglin (2018; which shows polar / subtropical jet separation in different seasons) and Manney et al (2021; which, while focusing on ENSO relationships, has several figures looking at subtropical / polar jet relationships) would be good references to cite along with Maher et al (2020) for showing seasonal variations in polar and subtropical jet stream locations and separation.

Lines 154–157, PO13 used a simple latitude criterion and demonstrated/stated that it didn’t work well. Manney et al (2014) did not distinguish but simply showed and discussed the spectrum of jets (I have no idea what you mean by that being “a more emergent form” and it doesn’t separate the jets into “groups”) – except for one climatological figure where they used a latitude criterion and discussed the fact that that was only useful for very broad climatological studies (consistent with PO13). If you are going to discuss this topic, you should note that Manney & Hegglin (2018) developed a physically-based identification of the subtropical jet and then defined the polar jets with respect to that – this is what the package originally developed by M11 has been using since then (eg, Manney et al, 2021, who discuss subtropical / polar jet relationships in the context of ENSO variations) (it would also be useful to have this method implemented in jsmetrics, though it does require tropopause altitude information, so would require reading additional fields, which at its current state, jsmetrics seems reluctant to do.) It might also be worth noting that some methods by their nature selectively identify the subtropical jet (e.g., though it isn't one of the ones currently implemented in jsmetrics, Maher et al., 2020), and that some (e.g., several papers by Winters and others (e.g., Winters et al, 2020, DOI: 10.1175/MWR-D-19-0353.1, and references therein) based on the method described by Christenson et al (2017, https://doi.org/10.1175/JCLI-D-16-0565.1) distinguish them by postulating that they occupy different altitude ranges (in a sense, a refinement of the common procedure of using low level wind maxima to identify the polar jet, but one that can be used while still capturing the level of the core of that jet, and in a similar spirit to Koch et al.).

Lines 190–191, I was unable to find this, though I believe I looked through the website of Keel (2023) and the associated GitHub project (except for the broken link mentioned in the general comments) thoroughly. This might have been helpful in assessing more definitively whether the implementation contained the fundamental aspects of the method.

Lines 204–206, the "details_for_all_metrics.py" file contains nothing in the "description" field for the vast majority of the methods (and the one I saw that was not empty did not give a description of the method). In the entry for M11, it makes it sound like only levels between 400 and 100hPa are needed to implement it, whereas (see general comment (3)) a deeper vertical range is required.

Lines 212–213, being able to see the pseudocode would have been very helpful in not only assessing to what degree the implementation matched the original (as mentioned above), but also in understanding what the implementation actually did in general.

Lines 225–227, as already noted, I found the manner in which the "implementation-level detail" was hidden made the package quite the opposite of user-friendly. Also, the implementation is critical for the user to understand where it means that the method may differ conceptually from that in the original paper. The user needs to know what each method, as implemented in jsmetrics, actually does, so if the implementation is hidden there has to be very thorough and complete documentation of that.

Line 237–239, by "7 metrics available in jsmetrics" it appears you mean the ones you identify as "statistics". Since most (if not all) of the "jet core algorithms" provide (at least in their original implementations) wind speed (a more accurate measure of "...speed of the jet stream…" than maxima in the u-component) it would help explain why you group these together if you note that these are all applied to the lower troposphere.

Line 241, you need to explain what the "violin plots" represent, either here or in the caption for the figure where they are first shown, in detail – that is, what the width and width variations mean, what the thin and thick lines along the centre mean, and what the (barely visible) white dot in the centre of the thick line means. Not everyone is familiar with violin plots and they are not obvious – you have to orient the reader so they can make sense of what you subsequently say based on interpretation of those plots.

Lines 242–243 and Figs. 2 & 3, I see no reason to show these defined inconsistently since you are not comparing them with the original papers here (which would be the only reason not to use consistent definitions). It would make much more sense to show these only once, as they are shown in Figs. 5 & 6, so that they can actually be compared with each other. This would also eliminate the problem of not showing some of them in some regions.

Lines 265–266, also, these lower tropospheric altitudes are far below the altitude of the core (maximum winds) of these jet streams, and regional averages would be lower than at an individual longitude grid point.

Lines 268–269, you should define what you mean by “structural uncertainties”.

Line 278, these five pressure levels represent vertical spacing / resolution that is inadequate to resolve upper tropospheric jet cores – see, e.g., comparison in Manney et al, 2017, of calculations using the M11 method on reanalysis model levels with the same reanalysis on standard pressure levels (which are poorer resolution than those in the models, but better than those used here).

Figure 4, why don’t you show the jet core wind speeds for all those methods that calculate it (both M11 and PO13 methods do)? Also it would be helpful to look at the jet core altitude information from those methods that calculate it (at least the two just mentioned) and see how close that actually is to 250 hPa – this would tell you how different you’d expect a jet core identified by those methods at whatever the actual altitude of the jet core is to agree with the winds at 250hPa.

Line 282, I’m not sure how really unique they are given that M11 is an extension of Koch et al and PO13 is nearly identical to M11.

Lines 287– 292, this does not seem like a particularly clear or even accurate description. M11 simply identify all (well, up to five in each hemisphere) jet cores at each longitude. They do not assess splitting (or merging), which I think(?) is what you are trying to get at in line 290 – note, however, that a faithful (but independent) implementation of the method by Homeyer & Bowman, 2013 does add a post-hoc algorithm such as you mention to “string the jets together” in longitude and identify splitting and merging. (A complete implementation of the current M11-based method would also provide subtropical / polar jet information.)

Lines 293–296. “Surrounding 8 grid cells” in what space (latitude / longitude / altitude?). If restricted to latitude / altitude, M11 also do this, and most certainly “make the assumption that the centres of the jet streams are important features in their own right” – that is, in fact, the central motivation for their method. (This would also hold for PO13.)

Line 298, I don't think it is intuitive that the differences would be amplified when aggregating into coarser time resolutions. In fact my first thought was that differences might be in some sense averaged or filtered out. Why do you expect this?

Lines 306–307, parametric uncertainty testing is a fundamental part of development of any algorithm, so some of this is typically mentioned in the original papers.

Lines 308–309, Manney et al (2017) focused on this; in addition, Manney & Hegglin (2018), and Manney et al (2021), and PO13 all used multiple input datasets (mostly reanalyses in these cases).

Lines 310–314, See comment on Figs. 2 & 3 – having both is redundant, and Figs. 5 & 6 provide far more useful information.

Lines 323, 326, and 340, by "temporal aggregations", do you mean you are averaging over that number of days before doing the jet identification / characterisation? (It looks like that to me since if you plot the points for all of those days / times of day I'd expect to see more points on the plot for each successively larger time interval.) If by "temporal aggregations" you mean "time mean", then just say "time mean". (Also, assuming that's the case, in the Fig. 7 caption replace "time scales" with something like "averaging periods".)

Lines 330–332, I'd expect more *transient* features to be "washed out" for longer averaging periods and more *persistent* features to remain better defined – what you are saying sounds contrary to this.

Lines 343–344 and 346–348, most (all?) of the "jet core algorithms" are also "purpose built" for extracting jet latitudes as one of their primary products (and jet wind speed as another) – e.g., these were a focus of results in PO13 and in the Manney-led papers from 2014 on. Why do you not just use the jet latitude and wind speed that those algorithms provide? It doesn't make sense to me to compare algorithms by comparing something obtained via a method they didn't use.

I have little to say in terms of specifics on Sections 5 and 6 – since they are very general discussion I think the related issues (such as the lack of success of the package as it currently stands, particularly with the current state of the documentation, in being "simple to use").

Minor / Technical points (typos, grammar, wording, etc; wasn't particularly reading for this, so just what I happened to notice):

Throughout, "ERA-5" needs to be replaced with "ERA5" as that acronym does not have a hyphen in it.

Line 1, I see no need to say “this planet’s” – so far as I know the jet streams are complex on all the planets (eg, Earth, Venus, Jupiter, Saturn, possibly others) we’ve studied them one (and given enough data, similar methods might be used for other planets)!

Line 11, I think you could and should leave out “or reduces” (“highlights” by itself covers that by implication). Also “exist” should be “exists” and “characterise” should be “characterises” (refers to “Each”, which is considered singular).

Line 23, “most popular” is a subjective assessment, you could just say something like “Among the most commonly used approaches is to develop…”

LInes 46–47, should be something like “...where there are extreme temperature and vertical pressure gradients…” (and I don’t think “vertical” is needed here).

Line 68, delete comma after “detect”.

LIne 71, I think this should be plural possessive, thus “jet streams’ “

Line 89, delete comma after “and”

Lines 109–110, It would be clearer (because this is the point that Manney & Hegglin specifically discussed) to move the Manney & Hegglin reference to immediately follow “synoptic-scale events”.

Line 115, “propagations” should be “propagation”

Line 116, wording not clear, do you mean “...are ‘jet streams’ such as those driven by eddy- or thermal processes…” Also doesn’t “driven by eddy or thermal processes” pretty much cover all ‘jet streams’, thus not really needed? (Unless you mean to say something like “...are ‘jet streams’ nor do they diagnose the eddy or thermal processes driving them…”)

Lines 179–181, I’m not sure this is even worth saying, since it would be unacceptable for any package aimed at community use not to have this flexibility.

LIne 237, should note that it is lower tropospheric u-component data that are used here.

Line 238, need serial (Oxford) comma after “North Pacific”

Line 255, “these metric” should be “these metrics” and “which” should be “that”.

Line 283, “which” should be “that”

Lines 285, 288, 289, probably other places, in this usage “jet-cores” should be “jet cores” (two words, not hyphenated).

Line 301, delete comma after “calculation”.

Line 322, at least add “e.g.,” before these two references (and / or “and references therein” after) since there are many, many papers that discuss this.

Line 353, do you mean that all the jet core algorithms use a wind speed threshold (in which case there should be a comma between “algorithms” and “which”) or that it includes none of the ones that do use a windspeed threshold (in which case “which” should be “that”)?

Line 381, should be “...similar metrics, e.g., for calculating…” (unless you are saying the following is an exhaustive list, in which case “i.e.” is appropriate but the commas are still needed).

Line 382, 394, 396 “which” should be “that”.