In `A diffusion-based kernel density estimator (diffKDE, version 1) with optimal bandwidth approximation for the analysis of data in geoscience and ecological research,' Pelz et al. introduce a diffusion-based approach to kernel density estimation, as well as its implementation in Python. From what I understand, they treat input data points as delta function sources in a diffusion equation and run the diffusion calculation forward until a given stopping 'time'; they also allow for a user-specifiable diffusivity, that is a function of data space, allowing higher effective resolution in certain areas of the data space. They show that the KDE estimation method performs comparably to other KDE implementations in the Python scipy package, and they show that it tends to have lower mean-squared error for relatively small (O(100)) sample sizes.

This paper presents a novel statistical model, with the apparent innovations in the paper being twofold: (1) the use of a spatially-varying diffusivity in a diffusion-based KDE estimator, and (2) implementation of the method in an open-source language. With respect to the journal, GMD, the paper presents a new method for statistical modeling in general, with applicability to geoscience domains; so it seems to be within scope of the journal (though I have further comments on that below). Based on this, the paper seems publishable in principle within this journal. Overall, this seems like a cool technique!

In its current form, the paper has a number of issues that lead me to recommend a major revision: (1) the paper has widespread use of notation that is not well-explained and is not likely to be widely-known by a geoscience audience; (2) the literature review and discussion misses some key, current literature that is highly relevant; and (3) the current form of the paper appears to advance a statistical method and only briefly applies the technique to geoscientific data, making it unclear whether this paper is really within the scope of GMD. These issues are described in more detail below.

Major comments

Widespread use of unclear notation

One of the biggest issues with this paper is the widespread use of jargon and notation that is not common in the geosciences. For example, the paper makes widespread use of set notation (e.g., line 75, Equation 2), which is not commonly taught in geoscience curricula (I have no idea what the $\hat{f}\mathscr{R} \times \mathscr{R}\ldots$ notation in Equation 2 means, and I have quite a bit of training in math and statistics, including having published in statistics journals myself!). In another example, := notation is used (e.g., line 225), and I am not sure what it means here; I'm used to reading it as `is distributed as', but I don't think that's what is meant here (and what does the $=:$ mean in equation 30?). If notation like this is to be used, I suggest defining what the notation means at first use. My main concern here is that the notation may end up being a barrier to people reading this paper, which would then also inhibit this paper's potential impact.

Somewhat related to the above, there is ambiguous usage of the symbol t. In some places it is clear that it is meant as the bandwidth (e.g., line 113). In others (e.g., the left-hand side of Equation 10), it seems to mean 'time' in the sense of time-evolution of the diffused quantity u. I don't think that the bandwidth and the time are meant to be taken as being equivalent in this paper, so a symbol other than t really should be used for the bandwidth. (Especially for this audience, where t is almost always reserved to refer to time.) That said, I'm genuinely confused about Algorithm 1, where lines 4 and 5 of the algorithm seem to clearly be setting T_p and T_f as bandwidths, but line 7 seems to be treating t as time and T_p as a maximum time; so maybe bandwidth and time really are the same in this paper? If so, that needs to be stated *very* clearly and perhaps repeatedly, since that's quite unintuitive.

In the end, I'm not certain I understood the method well enough to review it thoroughly, which is a major concern.

KDE literature review

The introduction does a good job of describing KDE and the background of KDE, giving references to bandwidth choice up through Scott, 2012. However, there are some recent innovations in kernel density estimation that are highly relevant here:

- Bernacchia and Pigolloti (2011) derive a method for choosing both the kernel and its bandwidth in an `optimal' way. Note also that there are several geoscience applications of this approach, which can be found by looking at literature citing this paper.
- Davies and Baddeley (2017) describe a spatially adaptive bandwidth approach, which seems relevant here given the spatially-adaptive diffusivity used in this paper
- Chacón and Duong (2018) overview KDE methods and the variety of bandwidth selection methods used; this is especially relevant since Duong authored the widely used `ks` package in R, which would be a good, additional state-of-the-science package to compare against in Figures 6, 7, 9.
- Another relatively recent book overviews KDE methods: https://link.springer.com/book/10.1007/978-3-319-71688-6

I would also add that, for this audience, the introduction would be well-served by listing a number of examples of the use of KDE in geoscience literature.

Clarifying how this manuscript is in scope for GMD

My final concern relates to the scope. In my initial thinking, I considered recommending rejection of this paper because it seems like it might be out of scope for GMD. Most of the content and innovation in the paper relates to a general-purpose statistical method, so it almost seems like this might be more suitable for a statistics journal like JABES. I think that part of this impression comes across in the emphasis on the development of the statistical method itself, rather than its application (and/or potential application) in geosciences. I think this partly relates to the technical notation comment above too; because the notation does not seem typical of a geoscience paper, it kind of feels like it was written for a stats audience.

I spoke with GMD editors about this, and they indicated that they do think it is potentially in scope. But the paper should make a clear case for how this new method advances geoscientific modeling, and I think it might be good to emphasize applications a bit more than the method itself. I also think that the discussion of the method (i.e., section 2) could be revised to be much less rigorous (which would also help with the statistical notation comment); focus on making sure that a large portion of your potential geoscience audience can understand this method (and your innovation to it) rather than focusing on precise mathematical language. It also might help to modify the introduction by mentioning some specific uses of KDE in geoscience papers and also modifying the discussion to relate the advances in this paper back to those papers; what might have been different/better about the past papers if they had used your method?

I would also add that the paper should be more explicit about the innovations of this specific paper: in the abstract, the introduction, in the method section, and again in the conclusions. I do see that the innovation of this paper is discussed in lines 544-549, which is great, but the way this is written, it is not clear what specific aspects of this paper are new relative to the citations mentioned. It might help to revise Section 2 such that it is clear which equations are essentially 'background information' and which equations (or which parts of the equations) contain the innovation in this paper.