How far can the error estimation problem in data assimilation be closed by collocated data?

by A. Vogel and R. Ménard

Reviewer's Report by R. Todling

Recommendation: Accept with revisions.

The present work revisits the problem of estimating relevant statistical information for data assimilation by employing residual-based collocation methods. The work presents a generalization of three-cornered-hat (3CH) and traditional collocation methods establishing precise statements about how many relevant statistics can be inferred from a given number of datasets that include different estimates of sought out quantities. The work also provides for an understanding of what one can expect to estimate given various dependencies among differing datasets. The work is full of insight and provides illustration from idealized settings.

I my view the work is sound, mathematically meaningful and represents an important contribution to the field. I recommend some revision in the text, mostly minor points. I do have a couple of broad comments which are presented below before the minor points.

Main concerns:

A My first main concern refers to the wide use of the word *innovation*. Although I understand the main motivation behind the work is data assimilation applications, the framework in the present article is general - it deals with second moment statistics of variables regardless of the context in which these appear. The difference fields appearing in equations such as (4) are what would be better referred to as *residuals*. I strongly suggest replacement of the word innovation with residual. Indeed, unless commenting on related works truly using innovations (e.g., Tandeo et al. 2020; Todling et al 2022; and others), most of the time the authors can omit either of the words; especially once stated initially that the covariances and cross-covariances dealt with in the work are really *residual* covariances and cross-covariances.

B Another issue for me relates to notation. It starts around line 139, when the authors introduce eq. (4). I understand that subscripts such as $i-j$ represent differences (residuals) derived from estimates $\mathbf{x}_i$ and $\mathbf{x}_j$ for datasets i and j respectively. It is never said that in such case, i must never equal j, as it would not make sense to calculate residuals of a dataset against itself. An alternative notation for the subscripts of $\mathbf{\Gamma}$ would be $i,j;k,l$ - in this, the pairs being use to calculate the difference vectors making up $\mathbf{\Gamma}$ are separated by a semicolon. This notation would also be more consistent with the notation in eq. (5), when the truth is introduced and the matrix represents an error covariance.

C I believe that in the considerations in section 3, and specially section 4, a relevant possibility for how possibly to get the precise estimates when dependence exists among the datasets has been overlooked. The others talk a lot about what happens

when the dataset are truly independent, or when there is dependence. But never really point out the important case when the dependent contribution in, say, eq. (26) vanishes as a whole. That is when, the datasets are such that

$$\mathbf{D}_{\tilde{i};\tilde{j}} + \mathbf{D}_{\tilde{k};\tilde{i}} - \mathbf{D}_{\tilde{j};\tilde{k}} = \mathbf{0}$$

The above is at the core of the Todling et al. (2022) findings. That is when the three datasets, (i,j,k) here, are connected in some very special (particular) way, that is, through the DA system, (i.e., these being the analysis, background and observation). I believe the possibility of finding special combination of datasets (for which the above holds) should be discussed in your work. Clearly, datasets that combine in such particular way are rather rare.

Minor points:

1. I wonder about the title a little bit. The work here is very general, I know data assimilation is the primary motivation for the application of the method(s) discussed and the work done in this work. But the fact is that the technique here applies generally, and independently of DA. Perhaps a better title could be: "When do collocated data provide for a closed error estimation problem?"

2. l. 24: "arises the question if" should read "raises the question whether".

3. With extreme respect to the authors, I recommend a close revision of the writing itself. I find use of very uncommon English words, which although not incorrect, seem rather usual, e.g., exemplary, approximative; there are also a number of articles, and other wording in the paper that could benefit from some attention. I try to point out some of these in what follows, but I only show so much. I can anticipate that most of the time a word like "exemplary" is better read as "example", and "approximative" as "approximate".

4. l. 30: "since decades" should read "for decades".

5. . 30: "recently be exploited" should read "recently been exploited".

6. l. 31: The works of Nielsen et al. (2022) and Todling et al. (2022) were done concurrently, basically with either being unaware of the details of the other. I believe your statement here would more fairly read "Nielsen et al. (2022) and Todling et al. (2022) were the first to independently use the generalized ...".

7. ll 33-35: I think the authors need to rephrase what comes after "However". A better sentence would perhaps be: "... framework. Indeed, Todling et al. (2022) shows that when the corners of G3CH are identified with the observation, background and analysis of variational assimilation procedures, only under the assumptions of optimality does the method obtains closed estimates for the three corners; in general, the problem cannot be closed." Notice this comment goes along comment C made above.

8. The "dot" notation used in eq. (4), and many others, has not been introduced. I authors should state that
$$\mathbf{x} \cdot \mathbf{y} = \mathbf{x}\mathbf{y}^T$$

9. Eqs. (6) and (7) do not require the term exposed in their second equalities.

10. Eq (6): I confess the notation of using superscripts in the equality signs in various equations is new to me. I have mixed feelings about it, but regardless of my feelings, the authors should explain what these are after they first appear in eqs. (6) and (7). That is, somewhere it should be stated that "superscripts and subscripts in the equality signs indicate what other equations were used to arrive at the given result".

11. l. 47: "since decades" should read "for decades".

12. l. 53: "additionally" should read "additional".

13. l. 70: word "approximative" would better read "approximate". The word approximative appears numerous times, I believe all instances would read better as "approximate" instead.

14. l. 76: word "exemplary" should be removed in this case - without loss of clarity.

15. l. 79: "requiring the knowledge" would better read "requiring knowledge".

16. ll. 84-85: "analyses or any" would better read "analysis and any".

17. l. 138: there needs to be an explanation (definition) for the meaning of the subscript notation with the standing up bar, as in $i|r$, i given r? Why do you need this notation here when it is not used anywhere else in the article?

18. Eq. (4): I find it somewhat unnecessary to have the notation include the points (p, q) explicitly. Given that $\mathbf{x}$ is a vector quantity the (p, q) indexes can be implicitly understood. In fact, most of your eqs. do not carry them.

19. ll. 173-174: This sentence should be moved to the definition statements made around eq. (4).

20. ll. 190-193: This would better read: "Thus, the covariance of any two datasets consists of the sum of the independent covariances associated with each dataset minus the error dependency covariance; this latter corresponding to the sum of the error covariances associated with each dataset, eq. (16)."

21. l. 240: "formulated as sum" should read "formulated as a sum".

22. paragr. ll. 243-247: you might want to add here that all the works mentioned in this paragraph associate what the authors call "dependent contribution" with the *cross-covariances of the random errors.*

23. l. 243: please replace "by Eq. (26)" with "in Eq. (26)".

24. ll. 260-261: There are lots of instances of the word "formulation" in these two sentences; the author might want to work on the text.

25. ll. 278-279: This sentence is very confusing. I think I understand what the authors mean, but I suggest rephrasing.

26. ll. 306-308: I believe the authors want to say that the "*independence* assumption *resembles* the innovation consistency *statement* of data assimilation, where the innovation covariance ..." —- notice that here, this is one of those places where the word *innovation* can and should be used.

27. l. 314: word "neglectable" should read "negligible".

28. l. 325: word "between" should be replaced with "among". Please notice there are other instances of "between all three" that should be revised accordingly.

29. l. 337: "a error" should read "an error"

30. l. 339: "allow a comparison" should read "allow for a comparison".

31. l. 396: " Eq. (32) is follows" should read " Eq. (32) follows".

32. l. 418: spell: "beeing".

33. l. 461: 'An discussion" should read "A discussion".

34. l. 464: "provides an exemplary demonstration" would better read "provides some demonstration".

35. l. 485: word "calculated" is not needed.

36. l. 495: "exemplary demonstrated" should better read "illustrated".

37. l. 511: "does only affect" should read "affect only".

38. Fig. 2a: why are the errors (bottom row) so diagonally dominant? Shouldn't these bottom panels be more like random patterns everywhere? Why aren't the errors in the diagonal of the order of the off-diagonal terms?

39. Figs. 2b and 3: why are the errors (bottom row plots) so asymmetrically dominant?

40. l. 515: "it's" should read "its".

41. l. 516: "requirement of assumptions" should better read simply "assumptions".

42. l. 522: "and 4.2 and is" should better read "and 4.2 is".

43. l. 523: please spell out "Apx".

44. l. 530: “solution of the problem” reads better as “solution to the problem”.

45. l. 533: word “whoever” should be removed.

46. l. 533: “came up with a too strong requirement” would better read “came up with an unnecessarily strong requirement that *ldots*”

47. l. 539: “estimates” should be in the singular.

48. l. 575: duplicate “of the”.

49. p. 24, conclusions: in regards to your last two paragraphs, and the generality of the method as you propose here, can you comment on the viability of the method to be used for, say, deriving estimates of background errors by using a combination of background fields from multiple DA systems. For example, suppose we collect 6-hour forecasts from IFS, GFS, CMC, GMAO, US Navy, etc — there is some dependency among all these datasets since for most part the background fields (short-range forecasts) are based on the assimilation of similar observations in all these systems — do you think your method would be able to infer reliable and perhaps better forecast error estimates than what we typically get from the NMC or ensemble methods? The same question can be made wrt analysis errors. Can you comment on this - if not in the paper, at least here to this reviewer.