

Report on ‘Quantifying the model structural error in carbon cycle data assimilation systems’, by Kuppel, Chevallier and Peylin

By Ian Enting.

This is a significant piece of work and appropriate for publication. I have some questions about the methodology, a number of suggestions regarding notation for making the work more accessible and have noted some minor errors.

Questions about methodology

- The assumption that error is stationarity in time (i.e. characterised by a single lag time) is questionable for a system dominated by strong seasonal variation. The authors should at least acknowledge (and maybe briefly discuss) the limitations/implications of this assumption.
- P2271, L 21: As one of the authors of the Kaminski et al paper, I don’t understand the claim of ‘equivalence’. Kaminski et al are discussing a truncation of the ‘model space’, while this paper is based on a truncated observation space.

Presentation of notation

This is a demanding body of work, and to help the reader I would suggest additional assistance with the notation involved in the estimation.

- A table of notation would seem to be appropriate. I give a few possible comments below (assuming that my interpretation of the paper is correct).
- In the figure captions, giving the mathematical expression of what is being plotted (as well as the verbal description) would be helpful.
- My understanding is that the b superscript, used for the ‘prior’, comes from Desroziers et al. where it refers to ‘background’ (i.e. the forecast from the previous cycle). The authors should note this meaning, or maybe consider whether symbol this is appropriate given that the calculation is a ‘batch’ calibration of the model, rather than (cf CarbonTracker) an assimilation system where the ‘state’ is being progressively updated with time.
- P2264, L11: ‘prior residuals’ \rightarrow ‘residuals from the prior model’
- P2264, L11: more seriously, \mathbf{D} (as expressed by eqn (1)) is the covariance for the distribution of these residuals – the covariance of the actual set of residuals is (or can be) only an **estimate** of this distribution.

- P2264, L20: shouldn't this be: 'to directly derive an estimate of \mathbf{R} ' and then have eqn (3) as something like $\hat{\mathbf{R}}^{\text{empirical}} = \mathbf{F}$ (Clearly, cf P2269, L 12, the authors understand this distinction, but to help others, they need to be more thorough in distinguishing (unknown) distributional quantities from the estimates of such quantities.)
- P2264, L19: Shouldn't the parameters (that are to be optimized) be denoted \mathbf{x} . (with \mathbf{x}^b being the (prior) **estimate**.)
- Also, if the superscripts a, b, o are abbreviations for words 'analysis', 'background', 'observations' rather than mathematical variables, then these superscripts should be in an upright font to distinguish them from variables.
- Since the Desroziers paper is quite complex, and the present paper only uses one aspect of that analysis, it may be worth summarising (in Appendix A) the relevant relation(s) from Desroziers to show why \mathbf{F} gives an estimate of \mathbf{R} .

My understanding of the notation

\mathbf{R} Covariance matrix for observation error. Subscripts refer to contributions to \mathbf{R} , superscripts refer to contexts, "hat" refers to estimates.

Other comments

- P 2261, L 1. My understanding is that the information from the model equations is spread in space as well as time.
- P 2261, L2–5. This sentence is somewhat clumsy and might benefit from being restructured (and perhaps noting that the 'strong constraint' approach is a choice, not a necessity).

Minor points

- P 2261, L 6: replace 'Bayes theory' by 'Bayes theorem' OR 'Bayesian estimation'
- P 2261, L 25: parameters \rightarrow parameter
- P 2266, L14: 20121 \rightarrow 2012 (as date for Crisp et al paper)
- P 2273, L10: 'occult' \rightarrow 'fail to capture'
- P2273, L 23: 'all the more that' \rightarrow 'particularly since'

File kuppelrev.tex; typeset September 27, 2012