Geosci. Model Dev. Discuss.,
doi:10.5194/gmd-2016-148-RC3, 2016

Geoscientific Model Development Discussions

Interactive comment on "Fundamentals of Data Assimilation" *by* Peter Rayner et al.

Anonymous Referee #3

Received and published: 14 September 2016

The authors of this manuscript are experts in applying data assimilation methods to problems in the geosciences, especially in the area of interpreting trace gas measurements in terms of surface sources and sinks. In this manuscript they discuss several popular data assimilation methods – variational data assimilation, Kalman filters, and particle filters – in the context of Bayes theorem, showing how, at root, they are all essentially solving the Bayesian inference problem in different ways. They do this using a common notation, derived from the Ide et al (2003) paper, and discuss differences in the assumptions of each approach. An example of one such difference that is whether dynamical errors are considered for the transport model or not.

As will be seen below in my detailed comments, I had a problem with the notation used in this manuscript, especially with the use of "yˆt", or the "true measurement" to represent a measurement vector. I would prefer a notation in which the terms "truth" and "true" refer to quatities that have specific, objective real values, as opposed to

error distributions associated with them. In data assimilation and estimation theory in general, it is usually the estimates of these "true" quantities, which have errors and thus error distributions associated with them. Here, instead, we have a "true" measurement with an error distribution associated with it – a confusing concept to one familiar with the standard estimation theory notation and concepts. Perhaps this notation is coming direct from the Ide reference – if so, this manuscript would have been a good place to correct it and set forth a notation that attributes the error more correctly to the quantities the error is more appropriately placed on. More generally, this seems to be related to the dropping of the notation indicating which quantities are estimates versus those that are not. I don't think this distinction can be attributed to the difference between the Frequentist and Bayesian viewpoints – what is given here just seems wrong to me. The notation problems lead to problems in the way the underlying issues are thought about, in my opinion.

A second issue is the discussion of dynamical errors. The authors do discuss the issue in Section 5.5, but for much of the manuscript they stick with a notation that rolls together the transport model and the observation operator into a single function H(x). This conflation of the two error sources into a single term was established early in the atmospheric inversion literature, unfortunately, and has often not been un-conflated even today. In other fields, however, the two were never conflated, and a general discussion such as this should do its best to keep them separated – frequently, this has not been the case in this manuscript, as is reflected below in my detailed comments.

In general, this manuscript is a helpful discussion of the inter-relatedness of several different estimation methods used in the geosciences and would be worth publishing. I hope the authors will consider my comments as providing a path for improving the manuscript further. There are some notational issues that must be corrected before publication (e.g. the use of x vs. z for the state).

Detailed comments:

p2 L29: Shouldn't the probability of the union of two disjoint events be the product of the probability of each, not the sum?

p5 L12: You seem to be using "system variable" interchangeably with "state variable"? Might it not be clearer just to stick with "state variable"?

p5 L19: Equation (2) and the discussion in the text. Here, you seem to be using variable 'x' to represent the state of the system, though in your notation in Table 1, you say you are using 'z' for that. It seems that you need to say that your target variable 'x' is in fact the same as the state of the system 'z', so that you can reasonably write 'H(x)' instead of the 'H(z)' that you put forth in the _text_ (but not the notation) in Table 1 (i.e. "H Observation operator mapping model state onto observables"). Also, since we are on this topic, is "target variable" meant to be synonymous with "control variable" (that is, the vector of those parts of the system that can be controlled or manipulated to get the desired outcome), used often in the control theory literature? If so, it would have been better for Ide (and you here?) to reserve 'z' for that, and use 'x' for the state variable (consistent with that existing control theory literature). What's done is done, I suppose...

p5 L21: "the system state x" – again, inconsistent with what you have in Table 1, where the state is "z"

p5 L28: "a refined PDF for yˆt": This talk of a PDF for yˆt I think is mis-conceived. You state yourself that "We stress that the system state and the true value of the measured quantity are particular values. Our knowledge of them is instantiated in the PDFs for x and yt." This makes it clear that there is a distinction between the true value itself (which doesn't have a distribution, but rather a fixed, actual value) and our best estimate of the true value, which is in error and has an error distribution. The blurring of the line between the two, which you have built into your notation here, is particularly unfortunate when it comes to the measurement, which you call yˆt. One can imagine that the system has a true state zˆt which, when measured perfectly (without error) would yield

the measurement value corresponding to that true state. It would be a natural extention of the concept of a "truth" to refer to this quantity as the "true" measurement; i.e. yˆt = H(zˆt), for the case where H is assumed to be a perfect measurement operator. In the real world, the measurement process is not perfect, the actual measurement would have errors (reflecting errors in our knowledge of how to make a proper measurement) that would cause these real measurements to deviate from the ideal measurement, the perfect measurement, the "true" measurement. If the errors in this measurement process were gaussian, one could specify a gaussian measurement uncertainty on each flawed measurement y, and use that to quantify the error between these real measurements and the "true" measurement that would be obtained in the absense of measurement error. Instead, the authors choose to use the notation yˆt for the flawed, real-world measurement, rather than the perfect measurement H(xˆt) (obtained with a perfect H). This may not make much of a difference if we are always dealing with the difference between an actual, flawed measurement and the underlying value that it is attempting to measure, H(zˆt), but from a conceptual standpoint, it is placing the label "true" on the wrong quantity and seriously confusing the issue. Those formulations that keep estimates separate from the underlying objective reality place the distinction between "truth" and error-affected estimates correctly with their notation, I think; the notation used here, in contrast, confuses where the error should be placed.

I would be much happier if the authors made a distinction between the "true" underlying measurement H(zˆt) (where H is perfect), and an actual measurement of that quantity, possibly affected by random measurement errors: that quantity is usually called something else, "z" for example, to indicate that it is a measurement prone to all the errors an actual measurement might have. Don't put the label "true" on that.

p5 L29: "The idea of a measurement being improved by a model is surprising at first." This can still be the case, but it would reflect an improvement of an _estimate_ y of the measurement rather than the true measurement yˆt. When thought of in those terms, it is not surprising at all. Why it appears surprising here is that the authors have used

the notation "yˆt" for the measurement – with that notation, it does appear surprising that you can improve upon something that is already "true".

p7, L19-20, Since there are two variables being discussed, it is not clear which variable the uncertainties should be couched in terms of, in the last sentence.

p7 L25: It is not clear here whether "the quantity" that the covariance is being calculated for is the "quantity of interest" in item #1 of the second list above, or of a target variable. Following Table 1, it seems like we need to calculate the uncertainties in the target variables, x. Why are we interested in the PDF of some other variable, even if it is "of interest", if it does not factor into the estimation problem? My understanding of the assimilation problem, using the notation laid out in Table 1, is that the uncertainties tracked in the method are those for the targeted variables x. Those seem distinct from "the quantity of interest" discussed here. Why case the uncertainties back onto a variable that is not the target variable?

p8 L28-29: "Note that neither the measurement nor the true value are random variables, it is only our state of knowledge that introduces uncertainty." It is not clear what this means. One could think of the true measurement as having a single value, reflecting objective reality, and the measurement being a random variable, reflecting the uncertainty contained in the measurement/modeling process. Why could the measurement itself not be considered a random variable, in that case?

p9 L5: "difference between the simulated and true value" of the measurements: This may get at the root of the problem I was having above with the definition of yˆt. It seems that the notation "yˆt" is being used as the actual measurement, including any measurement noise or biases, rather than as that measurement that would be given by the measurement operator operating on the true state in the absense of any measurement noise or errors in the operator. I would suggest that this notation be changed to something else.

p9 L10-13: "We frequently shorthand this as the data uncertainty (or worse data error)

when it is usually dominated by the observation operator. The resulting PDF describes the difference we might expect between the simulated result of the observation operator and the measured value. Thus analysis of the residuals (observation − simulated quantity) can help test the assumed errors. This forms part of the diagnostics of data assimilation treated in Michalak and Chevallier (2016)."

I would agree with this statement if the observation operator includes only the error in going between the propagated state vector and the observation. If, however, it includes also the error in the propagated state vector (and thus error in the dynamical model), then it is confusing two sources of error that are best kept separate (as in the formulation of the Kalman filter). Confusion on this point is prevelent in our field, resulting in model-data mismatch uncertainties being inflated much more than is truly justified. I see that the authors go briefly into this issue below, but perhaps greater emphasis on this point would be justified.

p9 L20: You need a PDF for the model error, not the model.

p9 L26-31: You have shown here how dynamical errors may be considered in the context of one implementation (variational data assimilation). It might be worth mentioning another common implementation, sequential filters (like the Kalman filter): since the state is estimated repeatedly across short spans, the dynamical errors can be accounted for explicitly by inflating the estimate of the state covariance as the state is propagated forward by the model (this is in fact built into the standard Kalman filter development).

p9 L9: "are equivalent" – equivalent to what? Adding some commas in this sentence might help to make it clearer.

p10 L13: "The observation operator can also be absorbed into the generation of posterior PDFs". It is not clear on the surface what this means. Could you please be more specific/clear, so the reader does not have to consult the reference to understand what is being discussed?

p10 L22-23: "Second, we see that the only physical model involved is the forward observation operator. All the sophisticated machinery of assimilation is not fundamental to the problem although we need it to derive most of the summary statistics."

For time-dependent problems in which a dynamical model is used, this dynamical model would be a second physical model that should be involved (this is the case for most of our geostatistical applications). The fact that it often is not involved in the equations we write down is an error in the way we approach the problem (i.e. using a strong dynamical constraint instead of a weak one (in the variational approach) or using a Kalman filter with dynamical errors added explicitly). The lumping of dynamical errors together with errors in the observational operator is a gross approximation that results in conceptual errors of the sort made here in this statement.

p13 L3-4: "As we saw in section 4 we also calculate the posterior PDF for yt the measured quantity." An oblique reference to this was given at the very end of Section 4, but no calculation was given. Perhaps you should give an equation at the end of Section 4 showing how the posterior PDF for yt is calculated, to support this statement you make here.

p13 L13-14: "The largest computation in this method is usually the calculation of H which includes the response of every observation to every unknown. This may involve many runs of the forward model."

Again, here, you are addressing the specific case where you conflate the dynamical model and the observation operator into a single function H. This is a specific approximation to the general case, which keeps the two separate.

"Once completed H is the Green's function for the problem and instantiates the complete knowledge of the resolved dynamics." This is only the case if you are ignoring dynamical errors. As the authors themselves note earlier in the paper, a more sophisticated analysis would allow errors in these Greens functions due to dynamical errors. It is hard to accept that these Greens functions include "complete knowledge" if they

do not account for these dynamical errors.

p14 L3-4: "...in which the state of the system is continually adjusted...". Since the true state of the underlying system had a single trajectory through time (i.e. has some objective real value, rather than a probability distribution), what you are really referring to here is some estimate of the state of the system. I believe your language and notation should reflect that.

p14 L6-8: "For a hindcast we can counter this by expanding our set of unknowns to include not only the current state but the state for several timesteps into the past. This technique is known as the Kalman Smoother (Jazwinski, 1970)"

To be more specific, this technique of including a number of previous times in the state is referred to as a FIXED LAG Kalman filter. This should be mentioned, as there are at least a couple other flavors of Kalman smoothers (fixed point and fixed interval).

p14 L22-23: "It is perhaps unfortunate that many treatments of data assimilation start from the discussion of a least squares problem and thus hide many of the assumptions needed to get there." You are making an assumption yourself here – that those who are using the least squares method want to make detailed assumptions about the statistics for their problem. If all they care about is getting an unbiased estimate and minimizing the standard deviation of their errors (measurement and prior), irrespective of what the higher moments of the PDF might look like, then the least squares approach is consistent and works just fine.

p14 L25-26: "Minimizing J is a problem in the calculus of variations and so the methods are usually termed variational." What you say here is overly-specific and doesn't really capture the essence of the problem. Really, minimizing J is a minimization problem. There are many numerical methods for minimizing a cost functional, and most of them are not variational. One does not need to get into the calculus of variations to understand that if one goes down-gradient on a manifold, one will get closer to the minimum. Most of the standard minimization methods use this concept as their basis. True, the

calculus of variations allows one to calculate gradients in a computationally efficient manner, and those gradients can be used in gradient-based descent methods to do the minimization, but this does not make these descent methods "variational".

p15 L5-7: For the case we are using here, in which transport and measurement are conflated in H, HˆT is the adjoint.

p16 L6-8: Amonng the disadvantages of the EnKF, you might note that inflation is often added to the ensemble to prevent the spread of the ensemble members from collapsing. This is often done in an ad hoc manner. Thus, effectively, the dynamical noise that is added with physical meaning in the straight Kalman filter is replaced with an ad hoc inflation term that has lost its physical meaning.

p16 L13: please add "given in Section 6.5" after "We parallel the description of the Kalman Filter algorithm" to help the reader remember where this was

Corrections to grammar, punctuation, etc.:

p1 L8 : add a comma after 'debate'

p1 L9: the semi-colon should be a colon, I think

p1 L13: add a comma after "For example"

p2 L23: capitalise "P" in "Pg"

p2 L32: replace the comma with a semi-colon after "definition"

p3 L18: add a comma before "the calculation"

Table 1, line describing 'd': Since H acts on the state z, not the vector of target variables, this should read "y - H(z)"

Table 1, line describing 'R': First, the variables inside U() should not be subscripted, as they are now. Second, the quantity inside U() should be "y - H(zˆt)", for the same reason as above.

Table 1, line describing 'U(x)': I am familiar with this expressed as being the uncertainty of an _estimate_ of x around the true value of x, xˆt. Similarly for the definition of "x" up top, there is usually a distinction made between an estimate of a vector of target variables, and a simple listing of what those variables happen to be. You attribute this to the Frequentist view of the world and drop the distinction, but I think it is getting you in trouble here – perhaps you can get around this by mentioning some of this in the description of "U(x)" and what you are assuming in defining it this way. In other words, how do you answer if someone asks you what the difference is between a vector of target variables x and their true value, xˆt? Wouldn't the vector x be the vector of true values? If so, how can U(x) be defined, if the difference is always zero? If the vector of x is not the vector of true values of x (xˆt), then what is it a vector of? (If not of estimates, then of what?)

p5 L7: correct the Laplace citation (put all within parentheses?)

p5 L28: correct to "measured"

p6, L2: add a comma after "problem"

p7 L28: Another word besides "reticence" might be more appropriate. "reticence" means a hesitence to speak. It sounds like you want something reflecting a hesitance to use the Bayesian approach, or to trust it.

p8 L5: here I would use "system" rather than "state", since the state represents the underlying system, and it is the functioning of the system that we care about.

p8, last line: for clarity, I would suggest adding commas after the initial "That is" and after "true value". Adding a "rather" before "than that" would also help.

p9 L3-4: put the two references inside of parentheses.

p9 L4-6, sentence starting with "The PDF": some commas in here would help this read better.

p9 L7-8: "Absent such direct verification calculations like sensitivity analyses or ensemble experiments (e.g. Law et al., 1996) give incomplete guidance." The subject of this sentence appears to be missing. Please reword to clarify this.

p9 L18: add a comma after "perfect"

p9 L19: add a comma after "condition"

p9 L25: put the reference in ()

p10 L2: add a comma after "model" or remove the one after "distributions"

p10 L10: add a comma before "while"

p12 L18: add a comma before "meaning"

p12 L31: change "there" to "their" (or possibly "three"?)

p14 Eq (7): The capital H's should be italic here – no need to linearize yet.

p15 L27; p16 L3; p16 L10: "NKF" – do you mean "EnKF" here?

p16 L6: Capitalize to get "The biggest"

Interactive comment on Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-148, 2016.