Response to Referees' Comments

Peter Rayner

October 11, 2018

I thank Amy Braverman and an anonymous referee for their comments. Both have highlighted a series of problems of presentation which I have addressed in the revision. The reviews have also prompted me to read more widely in the statistical literature and realise that the paper is, as I suspected, an application of existing theory. I now point to this theory, and focus more on developing the example and some of the possibilities and problems that arise in biogeochemical applications. I have made a series of general and specific responses to reviewers' comments. I detail the general responses first and address specific concerns from each referee below. I have placed referee comments in Typewriter font and my responses in Roman.

General Comments

1. I have removed the appendix and replaced it with a refernce to (*MacKay*, 2003, Ch. 28). thanks to the anonymous reviewer for pointing this out.
2. I have expanded the description of the TRANSCOM inversion as requested by Dr. Braverman.
3. I have made the conditioning on the data explicit in the various PDFs.
4. I have made a careful pass through the manuscript to regularise the typography.

Amy Braverman

General comments: `In general I like this paper a lot. However, I find it extremely difficult to follow because of some type-os and much notation with which I am not familiar. The notation seems inconsistent in distinguishing between fixed quantities and random ones, and indicating where conditioning has taken place.` It is easy at this point to get side-tracked into a discussion of the interpretation of probability. I think that Dr. Braverman's real concern is the discussion of the TRANSCOM case where the target variables and data should be more clear. See general response 2. I agree that it should be clear when conditioning has taken place, see general response 3.

Page 3, lines 6 and 7: Please define the random variables x and Hi . In what sense is P (x|Hi) the conventional data assimilation problem"? I hav added some explanatory text to clarify this.

2. Page 3, line 9: To what linear model" are you referring? A linear transport model represented by Hi ? What do you mean by over enough of the relevant pdfs"? Or, do you mean over enough of the support of the random variable Hi ? I should have said a linear observation operator. I have corrected this and expanded the text. "support" is precisely the language I needed.

3. Page 3, line 12 and 13: Is Hi the same as Hi ? H1 , . . . , HN are defined here as Jacobian matrices corresponding to N different transport models ...with un- knowns defined by the multivariate Gaussian G(xb , B)...". Which unknowns? I am following the notation of *Rayner et al.* (2016) so that $\mathbf{H}_i$ is the linearised form of H_i. This distinction is unnecessary here so I have changed to the linearised form throughout. $\mathbf{x}$ are the continuous variables described in the text now added at the head of the section.

4. Page 3, line 15: For each Hi our problem is the simple linear Gaussian inver- sion..." What does this mean? What is it you are trying to solve for or infer? Is it the flux that gave rise to the observed concentrations? The problem is more general than fluxes and concentrations although that is a common example and one I use later. Again, I hope the explanatory text at the head of the section explains the meaning of the symbols.

5. Page 3, line 16: Most importantly for us P (xa |Hi) is Gaussian." Please define xa . Should it be xa ? These should be bold throughout following *Rayner et al.* (2016). I have corrected this. 4$\mathbf{x}^{\mathrm{a}}$ is the analysis or posterior.

6. Page 3, lines 16 and 17: P (x, Hi) appears to be a joint distribution of two quanti ties: the vector-valued x and the matrix-valued Hi . Its unclear from the notation whether Hi is a random matrix or a fixed matrix. (On line 21, Hi is treated as random.) My guess is that it is fixed since the right side of the equal sign ap- pears to show the pdf of just one variable; presumably x. Is i a vector or a scalar? Please define i , Ui , and Wi . I have added these definitions. I have also switched from using $\mathbf{H}_i$ as the variable in the PDF to i since it is the index into the set of observation operators which is the target variable.

The expression P (x, Hi) = Wi G(i , Ui) does not define a proper pdf unless Wi = 1 since the area under the pdf must equal one. A more precise definition of a mixture would be in terms of ran- P 1 with probability wi , PK dom variables: X = K i=1 Ai Xi , Ai = , i=1 wi = 0 otherwise 1, and Xi G(i , Ui). I don't agree with this. The expression represents the probability that i is the correct model and $\mathbf{x}$ the value of the continuous target variable. The normalisation requirement is defined by the integral over continuous target variables and sum over

models. That is expressed by the extra constraint in the next equation. I have made this more explicit by writing W_i as $P(H_i$.

7. Page 3, line 23: Either x should be bold, or not. Do not mix within the same equation. Also, the notation G(i , Ui)(x) seems is very confusing (to me, at least). Do you mean that x is an argument to the function G? Why not write G(x|i , Ui)? Done throughout and a definition has been added in Section 2.

8. Page 3, line 26: In this equation Hi is treated as a non-random quantity. Above in line 21 it was random. Have you conditioned on it? If so, this distribution should be written as a conditional distribution. If not, then Wi is a random variable, not a fixed weight. I have rewritten the equations to make the probabilities explicit and expanded the explanation at the start of Section 2.

9. Page 3, line 26: I cant check this equation because I cant follow the derivation in Appendix A. See below. It appears from reviewer 2's comments that the derivation is a standard result which I now quote.

10. Page 3, line 27: I think there is an extra v" at the end of this line. removed.

11. Page 4, line 2: I assume that xb has a prior distribution somewhere because it is being treated as both random and fixed in various places. What is the prior distribution? $\mathbf{x}^{\mathrm{b}}$ is the mean of the prior distribution for the target variable $\mathbf{x}$. The confusion here raises a general question on which I seek editorial guidance. I have relied heavily on the notation and explanations in *Rayner et al.* (2016) thus making the current paper less self-contained. Should I move away from that and define the notation locally?

12. Page 4, line 16: Type-o. Corrected.

13. Page 4, Footnote is missing. Removed.

Appendix A, through page 12

1. Page 12, line 18: K was defined in the main text as a normalizing constant. What is K(Hi) here? Do you mean P (Hi)?

2. Page 12, line 21: I am confused by this equation. G is a function that has an argument and parameters. What are the parameters and what are the arguments in this expression? The definitions from Section 2, lines 13 and 14 should be restated here and clarified as indicated earlier.

3. Page 12, line 23: Please define . Why is x in bold while dx is not?

4. Page 12, line 25: I find the use of Hi as both a Jacobian and an indicator of model identity to be very confusing. Why not let Hi be the Jacobian of model i, and introduce a model indicator, say , an integer-value random variable taking values in 1, 2, . . . , M , where M is the number of models? See general comment 1.

Reviewer 2

It is important to note that all the theory from Section 2 is conditional on the same data y being used in all of the inversions. Is this the case in the Gurney study? What can be done when y differs across models? Yes, the TRANSCOM inversion studies kept the PDFs for prior and data constant and changed only the observation operator. This is now noted at the start of Section 3. the formalism as expressed here would struggle with different dimensions for different **y** and this is now noted at the end of Section 6.

I disagree with the introduction of JIC. Isnt this just the marginal log-likelihood of the ith component? I think it should be described as such. Also it is well-known that the marginal log-likelihood penalises for complex models and the marginal likelihood is commonly employed in model selection. I agree and thank the reviewer for sending me back to the literature. I have expanded the introductory section and reduced the theoretical development accordingly and now use the conventional terminology. I am more concerned with model weighting than model selection here but agree that the marginal likelihood rather than its logarithm is the more useful quantity. The logarithm is still a more convenient tool for presentation however and I continue to use it for figures.

Are you sure that if we replace H i BH Ti + R with I we retrieve the BIC? The BIC plugs-in maximum likelihood estimates of x into the log-likelihood, while the marginal likelihood (which you label JIC) integrates out x, which is different. I think they are equivalent. The same thing happens with the conventional χ^2 for an inversion. this uses the MAP for the target variables but, in the linear Gaussian case, when one substitutes this value, one obtains an expression involving the innovations and uncertainty projected into observation space.

P3 L30: The comment We dont believe that the relative quality of two model depends on the amount of data used to compare then... is slightly out of place. There are many texts that show that asmyptotically these criteria hold (under some assumptions). If there is reason to believe that the assumptions are not valid (and the criteria are hence not valid) for the flux inversion problem discussed, then some other theoretical foundation for the use of a different criterion is needed. My point is pedagogical rather than theoretical. Modellers tend to use comparison with data as a measure of relative quality. This is quite reasonable but the relative quality of models doesn't diverge as we use more data to compare them. Recall also that this paper is primarily concerned with model weighting rather than selection. I regard it as pathological that the weighting collapses towards a single model as the amount of data increases. I have rewritten this point.

Related to the previous comment, the theory shows that the posterior flux is a weighted Gaussian mixture with weights Wi . What is the theoretical justifica- tion of using the

marginal log-likelihood to weight when combining (as I think is implied when using the expression JIC-weighted)? The Wi being degenerate is unfortunate but not a valid reason. The same comment holds for the cross- validation metrics. the Gaussian mixture is a result rather than an assumption. Part of the confusion was caused by an error of writing rather than calculation. I used the marginal likelihood throughout for weights but was incorrect in several places in calling this the JIC.

P8 L25: I cannot find a reasonable justification as to why one should add R and R sample . There are justifiable alternatives for example in spatial analysis one would fit a space-time model to the residuals with nugget and use this for the new R. This will be guaranteed to be non-singluar and positive definite. I think that adding these two matrices can lead to unforeseen consequences in other settings. I have now expanded this section and discuss in more detail the task of describing the model contribution to uncertainty. The reviewer's example is another interesting approach. This kind of fitting of spatial models is certainly desirable if we have enough data to do it. In this case we don't but the problem still needs a resolution albeit it mperfect. I agree that the addition of the sample covariance of residuals is ad hoc and I now describe it as an example, along with another of the sample covariance of a priori simulations.

Other Comments

P1 L24: I do not agree with the comment When structural uncertainty enters the problem only ensemble methods are available. Although less direct, cannot one introduce an additional unstructured residual term in the model and see if that dominates? I agree, the comment was too strong and I have moderated it.

Section 2 and Appendix A need work to make the notation consistent. Just some things I picked up: Hi should be bold everywhere The lack of conditioning on the data y in all equations makes it hard to distinguish between prior and posterior distributions. Also the use of K(H i) as prior is confusing since K is a normalising constant in Equation (6). The matrix B in Section 2 has become P in Appendix A. In Appendix A, the PDF G() takes three arguments, while in Section 2 it only takes 2. Appendix A needs to be cleaned up, there are expressions like mathbf appearing, is not bold, the subscript i could be apparent on the LHS but not on the RHS, etc. See also my comment on clarifying the maths in Section A further below.

P12: I found that the maths from Equation (A10) onwards becomes a bit obscure. The result is correct, however I think more details are needed. (some algebra deleted) However, as the author stated I also think this proof should be readily available in some textbook as it is just the marginal log-likelihood of a Gaussian density. I group all these comments together since I have dealt with them all by essentially quoting the standard result on marginal likelihood

and leaving out most of the algebra.

P3 L27: Instead of assumption that we must choose one I would instead say assumption that the sum of the probabilities of the models given the data equals one. Strictly speaking, unless you are using a Bayesian estimator you do not need to choose any specific one model. I think here we *are* using a Bayesian estimator. I was trying to make a different point that "none of the above" is not an option. I have reworded this.

P4 L7: The statement Eq.6 is also the same expression as the maximum likeli- hood estimate in Michalak et al. (2005) is inaccurate. Estimate of what? Eq.6 is the marginal likelihood of the data under the ith component. But the expressions *are* the same. I now think this is because they represent the same thing, the marginal likelihood for a hyperparameter. I mention this in passing now but want to keep the paper focused on the practical use of the machinery.

P5 L32: What is the difference between JIC/N and JIC if N is constant? Does scaling affect any of the results and conclusions? Not here since we only use one dataset. See the above response on relative model quality.

P7 L9: Why model seven? From Figure 1 it looks like Model 8 is the best model? Apologies, this was a numbering from 0 vs numbering from 1 problem, I have corrected the text.

References

MacKay, D. J. C., *Information Theory, Inference, and Learning Algorithms*, Cambridge University Press, available from http://www.inference.phy.cam.ac.uk/mackay/itila/, 2003.

Rayner, P., A. M. Michalak, and F. Chevallier, Fundamentals of data assimilation, *Geoscientific Model Development Discussions*, *2016*, 1–21, doi:10.5194/gmd-2016-148, 2016.

Data Assimilation using an Ensemble of Models: A hierarchical approach

Peter Rayner[1]

[1]School of Earth Sciences, University of Melbourne, Melbourne, Australia

Correspondence to: Peter Rayner (prayner@unimelb.edu.au)

Abstract. One characteristic of biogeochemical models is uncertainty about their formulation. Data assimilation should take this uncertainty into account. A common approach is to use an ensemble of models. We must assign probabilities not only to the parameters of the models but the models themselves. The method of hierarchical modelling allows us to calculate these probabilities. This paper describes the approach, develops the algebra for the most common case then applies it to the TRANSCOM intercomparison. We see that the discrimination among models is unrealistically strong, due to optimistic assumptions inherent in the underlying inversion. The weighted ensemble means and variances from the hierarchical approach are quite similar to the conventional values because the best model in the ensemble is also quite close to the ensemble mean. The approach can also be used for cross-validation in which some data is held back to test estimates obtained with the rest. We demonstrate this with a test of the TRANSCOM inversions holding back the airborne data. We see a slight decrease in the tropical sink and a notably different preferred order of models.

1 Introduction

Models of any interesting biogeochemical system are inexact. Either they cannot include all interesting processes, the governing equations of processes are not known exactly or computational resolution limits the accuracy of the solution. Throughout this series we stress that quantitative descriptions should be inherently statistical meaning they must include information on the probability of any quantity, either inferred or predicted. This requires us to describe the uncertainty introduced into any quantity by that of the model. Model uncertainty is of two forms, structural and parametric. Structural uncertainties occur when we do not know the functional forms that relate the inputs and outputs of the real system or that control its evolution. In biogeochemical models these functional forms are exactly specified so that uncertainty is usually manifest as an error. Parametric errors occur when the functional forms are well-known but there is uncertainty in various quantities such as constants in physical equations, initial values or boundary conditions. Uncertainties in model predictions arising from parametric uncertainty can be generated by semi-analytic error propagation (e.g. Scholze et al., 2007; Rayner et al., 2011) or by generating ensembles of model simulations from samples of the PDFs of parameters (e.g. Murphy et al., 2007; Bodman et al., 2013).

~~When structural uncertainty enters the problem only ensemble methods are available. These are usually codified as Model Intercomparison Projects (MIPs~~Ensemble methods dominate the study of model uncertainty. The most common approach is Model Intercomparison) of which the Coupled Model Intercomparison Project (Taylor et al., 2012) for the physical climate

and C^{4}MIP (Friedlingstein et al., 2006) for the global carbon cycle are prominent examples. The MIPs play a crucial but controversial role in quantifying uncertainty. First, they may underestimate uncertainty since it is impossible, even in principle, to know how well a given ensemble properly samples the manifold of possible models. On the other hand not all models are equally credible. They do more or less well at tests like fitting observations or conserving required quantities. This has led to the application of Bayesian Model Averaging (e.g. Murphy et al., 2007) in which models are tested against some criteria (such as fit to observations) and their predictions weighted accordingly.

Inverse problems or data assimilation as discussed in this volume generally treats parametric uncertainty. It uses observations and statistical inference to improve knowledge of the uncertain values (see Rayner et al., 2016, and references therein for a general introduction). Structural model uncertainty must still be included and indeed it often dominates other uncertainties. Model uncertainty is hard to characterize with analytic PDFs since errors in the functional forms will project systematically onto errors in simulated quantities. Hierarchical approaches (e.g. Cressie et al., 2009) provide a mechanism for including uncertainties over the choice of model into the formulation. For an ensemble of models this involves introducing an extra discreet variable (the index of the set of models) into the problem and calculating its probability. This probability goes under several names, ~~most commonly~~ e.g. the Bayes Factor (Kass and Raftery, 1995) or the evidence (MacKay, 2003, ch.28). We can then calculate probability distributions for model parameters as weighted averages over these model probabilities. Hence this application of hierarchical Bayesian modelling is closely related to Bayesian Model Averaging (Hoeting et al., 1999; Raftery et al., 2005).

Ensemble methods are rare for biogeochemical data assimilation since there are few problems for which a useful population of assimilation systems currently exists. The clearest exception to this is the case of global scale atmospheric inversions where the TRANSCOM intercomparison (Gurney et al., 2002, 2003, 2004; Baker et al., 2006) used an ensemble of atmospheric transport models and common inversion systems to infer regional CO_2 fluxes from atmospheric concentrations. All these studies faced the problem of estimating properties of the ensemble such as its mean and some measure of spread. Throughout they opted for the ensemble mean and two measures of spread, the standard deviation of the ~~MLE~~ maximum a posteriori (most likely) estimate from each ensemble member and the square-root of the mean of the posterior variances of the ensemble. This treated all members of the ensemble equally.

Equal weighting was challenged by Stephens et al. (2007) who compared the seasonality of vertical gradients in model simulations and observations. They found that only a subset of models produced an acceptable simulation and that this subset favoured larger tropical uptake than the ensemble mean. Pickett-Heaps et al. (2011) compared simulations using optimized fluxes with airborne profiles. This required running optimized fluxes through the forward model used to generate the Jacobians. Of the four models tested TM3 performed substantially better against this extra data than the other three.

Both the cited studies used data not included in the inversion, a procedure often called cross-validation. Cross-validation asks whether new data enhances or reduces our confidence in previous estimates while Bayesian model averaging calculates our relative confidence in two models. We shall see that the machinery needed to answer these two questions is very similar.

The outline of the paper is as follows. In Section 2 we ~~develop~~ review the necessary machinery~~although the detailed algebra is relegated to an appendix~~. Section 3 describes an application to the TRANSCOM case including an extension to treat covarying

model errors. Section 5 discusses the use of the machinery for assessing cross-validation. Section 7 compares the technique with other model evaluation methods as well as discussing some computational aspects.

2 Theory

~~In traditional data assimilation we do not have a choice over the model relating unknowns to observations~~The following can be regarded as a development of ideas described in (Jaynes and Bretthorst, 2003, Ch.21) or (MacKay, 2003, Ch.28). the standard data assimilation problem seeks to improve knowledge of some target variables in a model given observations. We express our knowledge as probability density functions (PDFs) and the mathematical operations are multiplications of PDFs for the prior knowledge of the target variables, the observations and the observation operator which relates the target variables to the observations. In most applications the target variables are continuous quantities such as model parameters, initial or boundary conditions. Following (Rayner et al., 2016)[Eqs. 2,3] we write

$$p(\mathbf{x}|\mathbf{y},H) \propto \int p(\mathbf{x}|\mathbf{x}^{\mathrm{b}}) \times p(\mathbf{y}^{\mathrm{t}}|\mathbf{y}) \times p(\mathbf{y}|H(\mathbf{x}))d\mathbf{y}^{\mathrm{t}} \tag{1}$$

where $\mathbf{x}$ represents the target variables, $\mathbf{y}$ the observations, the superscript b represents the background or prior value, the superscript t represents the true value and H represents the observation operator. the left-hand side of Equation 1 represents the probability distribution for the target variables given both prior knowledge and the observations. We add H to this left-hand side to emphasise that the PDF also depends on H.

In the usual case of data assimilation we only have one observation operator. Thus we often forget that the posterior PDFs for ~~unknowns~~ target variables are implicitly dependent on the ~~choice of model~~observation operator. Where an ensemble of ~~models~~ observation operators is available we ~~must make this choice explicit. The~~ can no longer assume certainty over which one we should use. The ith observation operator H_i becomes part of the target variables so instead of calculating $P(\mathbf{x}|\mathbf{y})$ we now seek $P(\mathbf{x}, H_i|\mathbf{y})$.[1] The hierarchical approach factorises this joint PDF of ~~models~~ observation operators and unknowns using an expression known variously as the chain rule of probabilities or the law of total probabilities

$$P(x, H_i) = P(x|H_i)P(H_i) \tag{2}$$

~~$P(x|H_i)$ is the conventional data assimilation problem so the new step is to find $P(H_i)$~~ Substituting Equation 1 into Equation 2 we see that the hierarchical and nonhierarchical PDFs differ only by the factor $P(H_i|\mathbf{y})$ and we hence need to calculate this term.

We will develop the theory for the simplest linear Gaussian case. Here many of the resulting integrals have analytic solutions. The approach will hold for nonlinear ~~models~~ observation operators provided they are approximately linear over ~~enough of the relevant PDFs~~the significantly positive parts of the PDF. The qualitative ranking of models is unlikely to be sensitive to weak nonlinearities since, as we shall see, the discrimination among models is strong.

[1] The true target variable is i, the index variable on the set of observation operators but we will continue to use H_i to make it clear to what this index refers.

We follow the notation of Rayner et al. (2016). Take a collection of linear ~~models~~ observation operators with Jacobians $\mathbf{H}_1 \dots \mathbf{H}_N$, with prior estimates of the unknowns defined by the multivariate Gaussian $G(\mathbf{x}^{\mathrm{b}}, \mathbf{B})$ and data defined by the multivariate Gaussian probability $G(\mathbf{y}, \mathbf{R})$.

For each $\mathbf{H}_i$ our problem is the ~~simple~~ linear Gaussian inversion described in (Rayner et al., 2016, Section 6.4). Most importantly for us ~~$P(x^{\mathrm{a}}|\mathbf{H}_i)$~~ the posterior PDF $P(\mathbf{x}|\mathbf{y}, \mathbf{H}_i)$ is Gaussian. Thus our posterior for the ensemble is a mixture distribution of Gaussians

$$P(\mathbf{x}, \mathbf{H}_i|\mathbf{y}) = W \propto P(\mathbf{H}_i|\mathbf{y}) \times G(\mu \mathbf{x}^{\mathrm{a}}_i, \mathbf{U}\mathbf{A}_i) \tag{3}$$

~~With the constraint that~~

$$\sum_i W_i = 1$$

~~The marginal probability for the models is obtained by integrating over x thus~~

$$P(\mathbf{H}_i) = W_i$$

~~and~~ where $\mathbf{x}^{\mathrm{a}}_i$ is the maximum aposteriori probability estimate or analysis for the ith Jacobian and $\mathbf{A}_i$ is the corresponding analysis covariance. The constant of proportionality is set such that $\sum_i P(\mathbf{H}_i|\mathbf{y}) = 1$. As usual with a joint PDF we obtain the marginal probability for ~~x by summing over models:~~

$$P(\mathbf{x}) = \sum_i W_i G(\mu_i, \mathbf{U}_i)(x)$$

~~Our problem then is to find W_i. Although the development is probably not novel we derive it in Appendix A for completeness; quoting the result:~~

$$W_i = K \left|\mathbf{R} + \mathbf{H}_i \mathbf{B} \mathbf{H}_i^T\right|^{-1/2} \exp\left[-\frac{1}{2}(\mathbf{y} - \mathbf{H}_i \mathbf{x}^{\mathrm{b}})^T \cdot (\mathbf{R} + \mathbf{H}_i \mathbf{B} \mathbf{H}_i^T)^{-1} \cdot (\mathbf{y} - \mathbf{H}_i \mathbf{x}^{\mathrm{b}})\right]$$

~~where the normalisation constant K is set such that $\sum_i W_i = 1$. v~~ a variable by integrating over all others. In the case of the set of observation operators this integral reduces to a sum. $P(\mathbf{H}_i|\mathbf{y})$ is the marginal likelihood for a Gaussian (Michalak et al., 2005, Eq.10)

$$P(\mathbf{H}_i|\mathbf{y}) = K \left|\mathbf{R} + \mathbf{H}_i \mathbf{B} \mathbf{H}_i^T\right|^{-1/2} \exp\left[-\frac{1}{2}(\mathbf{y} - \mathbf{H}_i \mathbf{x}^{\mathrm{b}})^T \cdot (\mathbf{R} + \mathbf{H}_i \mathbf{B} \mathbf{H}_i^T)^{-1} \cdot (\mathbf{y} - \mathbf{H}_i \mathbf{x}^{\mathrm{b}})\right] \tag{4}$$

2.1 Interpretation

Provided $\mathbf{x}^{\mathrm{b}}$ and $\mathbf{y}$ are independent, $\mathbf{R} + \mathbf{H}_i \mathbf{B} \mathbf{H}_i^T$ is the variance of the prior mismatch $\mathbf{y} - \mathbf{H}_i \mathbf{x}^{\mathrm{b}}$ (as noted by Michalak et al., 2005) so Eq. 4 represents the probability of simulating the observations given the prior estimate and related uncertainties. Quite reasonably, the higher this probability the more likely the model. We can say equivalently that the model performance should be

judged by the normalised prediction error (simulation − observation divided by its variance) penalised by the expected range of the predictions or the volume of the data space occupied by the prior model and its uncertainty (see discussion in MacKay, 2003, Ch.28).

Eq. 4 ~~is also the same expression as the MLE in Michalak et al. (2005) . That is reasonable, maximizing the relative probability of a model should mean maximising the Michalak MLE~~occurs in other hierarchical contexts such as the calculation of covariance parameters by Michalak et al. (2005) and Ganesan et al. (2014). This is understandable since the submodels in all three cases are the classical Gaussian problem. We note that these two papers used Eq. 4 to tune covariance parameters which may change the relative weighting of models. It raises the question that relative performance of models may depend strongly on whether the inversion is well-tuned for that model. The algorithm in Michalak et al. (2005) consists of tuning a scaling factor for prior covariances to maximize ~~W_i~~$P(\mathbf{H}_i)$ (though in their case there is only one model). We can test the sensitivity to a uniform scaling of $\mathbf{B}$ and $\mathbf{R}$ by a factor α. Increasing α increases the determinant so decreases the first ~~part of W_i~~ factor of $P(\mathbf{H}_i)$ while it decreases the negative exponent and so increases the second part. The balance is a relatively subtle change. ~~We will investigate later~~ In Section 3 we will investigate whether this is enough to change the ranking of models in one example.

The exponent in Eq. 4 is also the minimum value of the cost function usually minimised to solve such systems. It is often denoted $\frac{1}{2}\chi^2$. In a statistically consistent system χ^2 is equal to the number of observations (Tarantola, 1987, P.211). We often quote the normalized χ^2 as $\frac{\chi^2}{n}$].

Note also that for a given $\mathbf{B}$ and $\mathbf{R}$, Eq. 4 is extremely punishing on inconsistency. For example with $n = 10000$, a normalized χ^2 of 1.01 instead of 1 yields a ratio of probabilities for the two models of $e^{50} \approx 10^{21}$. This is unrealistic and is an example of the "curse of dimensionality" (Stordal et al., 2011) in which distances between points in high-dimensional spaces tend to infinity. We shall address one approach to resolving this problem ~~later~~in Section 4.

2.2 Relationship with Other Criteria

~~W_i~~ $P(\mathbf{H}_i)$ is related to several other measures of model quality. Define

$$\text{JIC} = -2\log(\frac{W_i}{K}\frac{P(\mathbf{H}_i|\mathbf{y})}{K}) = \log\left|\mathbf{H}_i\mathbf{B}\mathbf{H}_i^T + \mathbf{R}\right| + \chi^2 \tag{5}$$

which we will call Jaynes's Information Criterion after Edwin Jaynes. [2] The change of sign means smaller values of JIC correspond to more likely models. ~~We see that increasing uncertainty of either the prior or the data will decrease χ^2 but increase $\log\left|\mathbf{H}_i\mathbf{B}\mathbf{H}_i^T + \mathbf{R}\right|$.~~

The JIC is related to ~~Schwarz's Bayesian~~ other criteria for model selection such as the Akaike Information Criterion (Akaike, 1974) and Schwartz Information Criterion (also called the Bayesian Information Criterion,BIC) (Schwarz, 1978)~~which penalizes~~ . Both these criteria penalise models for adding parameters. ~~Instead of the term $\log\left|\mathbf{H}_i\mathbf{B}\mathbf{H}_i^T + \mathbf{R}\right|$ the BIC contains the number of parameters n. We note that if we replace $\mathbf{H}_i\mathbf{B}\mathbf{H}_i^T + \mathbf{R}$ with an identity matrix we obtain Schwarz's criterion. The BIC takes no~~ Neither take account of different prior uncertainties among parameters or different sensitivities of the observations to these parameters.

[2]

3 The TRANSCOM Example

The **TRANSCOM** III intercomparison (Gurney et al., 2002, 2004; Baker et al., 2006) ~~used a series of~~ was designed to investigate the impact of uncertainty in atmospheric transport models ~~, represented by matrices of Green's functions, to estimate sources and uncertainties~~ on the determination of CO_2 sources and sinks. The target variables were the mean CO_2 flux from each of 22 regions (11 land and 11 ocean) for the period 1992–1996. These fluxes excluded fossil fuel emissions and a data driven estimate based on ocean and atmosphere measurements (Takahashi et al., 1999). Prior estimates and uncertainties were gathered from consultation with experts in each domain. The data was the average CO_2 concentration from 77 stations. Participants in the intercomparison calculated Jacobians by inserting a unit flux into an atmospheric transport model corresponding to each region. there were 17 participating models so our space of target variables consists of 22 flux components and an indexed set of 17 models $\mathbf{H}_i$.

The inversions for the flux components are carried out by changing $\mathbf{H}$ with all other aspects held constant. The authors then created pooled estimates of ~~these quantities~~ the posterior fluxes such as the mean~~estimate~~, the mean uncertainty (averaging all the posterior uncertainties) and finally the "between model" spread, calculated as the covariance among the ~~individual mean estimates.~~

posterior fluxes for each model. In all these calculations we weighted every model equally. ~~An obvious objection is that not all models are equal and that we should weight models by some measure of their quality. Stephens et al. (2007) filtered the models according to whether they fitted some independent data of vertical profiles. Pickett-Heaps et al. (2011) also tested their source estimates against independent data. Neither paper, however, took account of the residual uncertainty in the sources. Even if a given inversion didn't fit a profile, was it possible to vary the sources within the uncertainties allowed by the inversion to still fit the profile along with the original surface data?~~

~~As a first test of the theory developed~~ What happens if we apply the methods described in Section 2 ~~we calculate the probability distribution for the models used in the **TRANSCOM** III Level 1 experiment (Gurney et al., 2002). This experiment inverted the annual mean distribution of CO_2 concentrations using 17 different atmospheric transport models.~~ to calculate pooled estimates?

Figure 1 shows ~~the~~ a slightly modified JIC for the seventeen models for the cases without (top) and with (middle) tuning following Michalak et al. (2005). The modification consists of displaying $\log_{10}$ rather than the natural logarithm. For the tuning cases we used one multiplier each for ~~the~~ $\mathbf{P}$ and $\mathbf{R}$. We see a large range of weights, 11 orders of magnitude for the untuned and 14 orders of magnitude for the tuned cases. This certainly reflects the "curse of dimensionality" mentioned earlier. For the same reason there is a strong focus of weight on a few models. Tuning intensifies this focus though it leaves the ranking almost unchanged. We conclude therefore that variation in model performance (as measured by the JIC) ~~do~~ does not reflect the quality of tuning of the inversion but something more fundamental about the models and data. Henceforth we consider only the untuned case.

~~Once we have calculated the Gaussian weights we have described the PDF of a joint manifold over the spaces of models and physical parameters. By calculating marginal probabilities from this PDF we can make statements about models or parameters~~

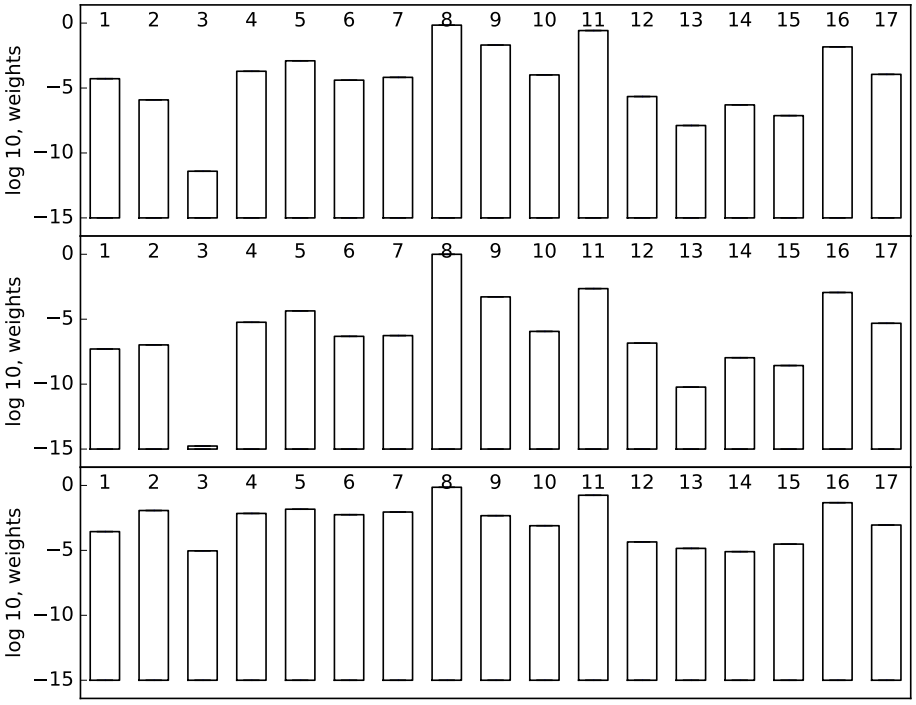

Figure 1. $\log_{10}$ of ~~w_i~~ $P(\mathbf{H}_i|\mathbf{y})$ for the untuned (top), tuned (middle) and case with residuals used for $\mathbf{R}$ (bottom) transcom inversions.

~~separately~~In the next two sections we consider the marginal probabilities to investigate the relative probabilities of different models and the pooled flux estimates.

3.1 Model Probabilities: Comparing Model Performance

The Gaussian weights derived in Section 2 are the probabilities that a given model is the correct one for matching the data under the ~~very strong~~ assumption that we must choose one ~~(the theory does not include a "reject all" option)~~(see Jaynes and Bretthorst, 2003, P136 We must, however, be careful not to over-interpret these probabilities as measures of model quality. In the first place, the JIC, like the BIC and χ^2 grows with the number of observations. So, then, does the divergence among models, an effect intensified when we take exponentials to calculate probabilities. We don't believe that the relative quality of two models depends on the amount of data used to compare them even if our ability to distinguish between them does increase as we add data. We can consider the normalised JIC JIC/N (where N is the number of observations) as a generalisation of the normalised χ^2. This ranges from a minimum of 0.01 to 0.67. The very low value should not be interpreted as representing an absolute quality of fit since we have normalised the probabilities to sum to 1. Rather it tells us that the apparently large change in the weights is a result of much smaller differences in the relative quality of the fit coupled to large amounts of data.

3.2 Ensemble Means and Variances

We can calculate various statistics of the ensemble using well-known properties of Gaussian mixtures. the mean is calculated as

$$\mu = \sum_i w P(\mathbf{H}_i \mu | \mathbf{y}) \mathbf{x}^a{}_i \tag{6}$$

Note that this collapses to the conventional mean if all weights are equal. the variance is calculated as

$$\sigma^2 \mathbf{A}^* = \sum_i w P(\mathbf{H}_i \sigma | \mathbf{y}) \left[\mathbf{A}^*{}_i{}^2 + w_i (\mu \mathbf{x}^a{}_i - \mu\mu)^2 \right] \tag{7}$$

~~We have used roman font for σ and μ to indicate they are scalars.~~ The Superscripts $*$ indicates we consider only the diagonal of the relevant matrices; Equation 7 only accounts for the variance not the covariance of the estimates. The second term in Equation 7 includes the spread of the means for each model. If all the ~~w_i~~ $P(\mathbf{H}_i|\mathbf{y})$ are equal, Equation 7 collapses to the "total uncertainty" metric used by Rayner (2004) to incorporate both the "within" and "between" model uncertainty described in Gurney et al. (2002).

Figure 2 shows the equally-weighted and ~~JIC-weighted~~ probability-weighted case for the **TRANSCOM** regions, in a format following Gurney et al. (2002). Here we do not show the "within" and "between" metrics separately since the Gaussian mixture naturally combines them. The focus of ~~w_i~~ $P(\mathbf{H}_i|\mathbf{y})$ on a few models (70% on one model) might suggest that the uncertainty in the weighted case should be far smaller than the equally weighted traditional case. Figure 2 shows this is not the case. Both the means and uncertainties for the two cases are quite similar.

The agreement of the means is explained by a result from Gurney et al. (2002). They noted that the mean simulation from their equally-weighted ensemble produces a better match to the data than any individual model . The ~~JIC-weighted~~ probability-weighted flux is constructed to maximize the posterior probability across the model ensemble and parameter PDFs thus its mean should also produce a good match. It is hence no surprise that the preferred model ~~seven~~ eight is the model closest to the unweighted model mean. Recalling that the ensemble weights this preferred model at 70% we see good agreement between weighted and unweighted means.

The similarity in the weighted and unweighted total uncertainty is partly a result of the weak data constraint in our problem. Gurney et al. (2002) noted that for almost all regions the "within" uncertainty was larger than the "between". Furthermore the posterior uncertainties produced by each model are rather similar so that the weighted and unweighted contributions in equation 7 are similar. The contributions of the "between" uncertainty are different in the weighted and unweighted cases but, since these are smaller than the other contribution, we do not see a large final difference. This would change in cases where the constraint afforded by the data (as evidenced by the uncertainty reduction cf the prior) was large.

4 Improved Treatment of Observational Covariance

Although mathematically correct, the strong discrimination among models ~~of~~ by the JIC is not intuitively reasonable. One reason for the strength of the discrimination is that each datapoint makes an independent contribution to the PDF. This is not an

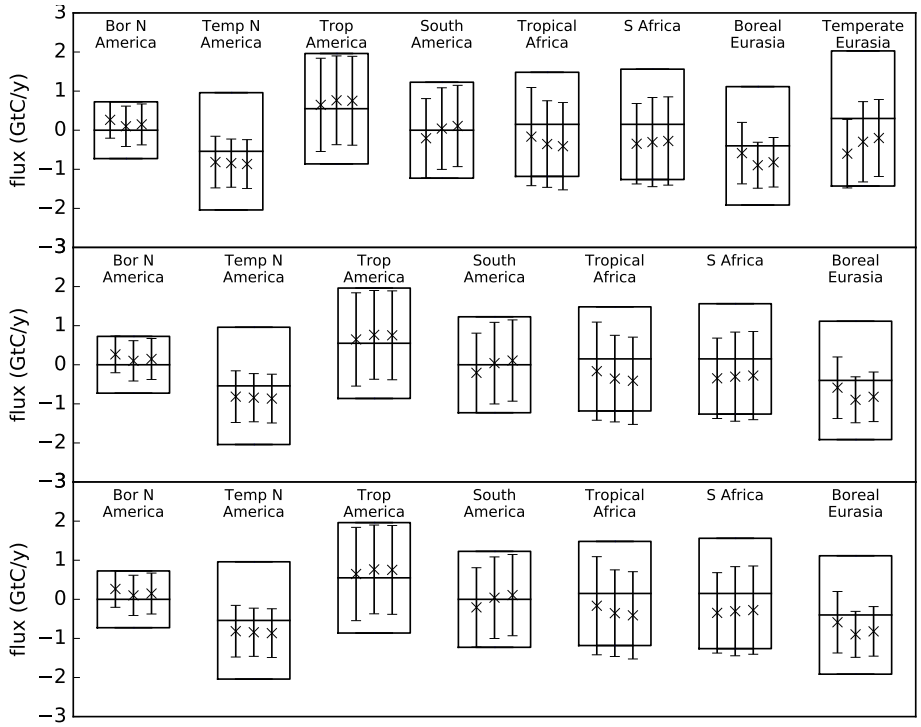

Figure 2. Prior and posterior uncertainties for regional fluxes from the TRANSCOM intercomparison following Gurney et al. (2002). The centre line of each box shows the prior estimate of the mean while the box limits show the $\pm 1\sigma$ uncertainties. The three bars show the mean (marked with "x") and $\pm 1\sigma$ uncertainty denoted by the length of the bar. The uncertainty is that of the ensemble including both the uncertainty for each model and the dispersion among model means. The left bar shows the equally weighted case, the middle bar the case for the ~~w_i~~ $P(\mathbf{H}_j|\mathbf{y})$ and the right bar the case with covariance of residuals included.

error in the formulation of the JIC but rather the PDF associated with the data in the underlying assimilation.[2] Physically this asumption says that if a model makes an error at one station, one cannot assume it will make a similar error at a nearby station. The physical coherence of atmospheric transport processes makes this most unlikely, even if subgrid heterogeneity lends some independence to the two stations.

There ~~is little machinery available for assigning, a priori, uncertainty correlations for the model . Lauvaux et al. (2009)~~ are two major approaches to characterising the model error covariance, either a priori or a posteriori. A priori we would like some machinery for calculating how uncertainties in model components or drivers project into model simulations. Lauvaux et al. (2009), for example, described a mechanism for calculating correlations in simulated tracer distributions due to correlated meteorological uncertainty but this is not a comprehensive description~~. Thus we are forced to fall back on analysis of the posterior residuals. This technique was previously employed by Kuppel et al. (2013) who derived various aspects~~ , i.e it leaves out many sources of uncertainty. If we have an ensemble of models we can use the ensemble of simulations using the prior value of the

[2]Strictly speaking it is the model PDF from Rayner et al. (2016), but we have combined model and data uncertainties following their Section 6.4

target variables as a measure of the model contribution to ~~the statistics of the model-data mismatch using techniques described by Desroziers et al. (2005)~~uncertainty. This was suggested by Tarantola (1987). We can write this as

$$\mathbf{R}_{i,j}^{\text{prior}} = \overline{(\mathbf{H}\mathbf{x}_i^{\text{b}} - \overline{\mathbf{H}(x)_i^{\text{b}}})(\mathbf{H}\mathbf{x}_j^{\text{b}} - \overline{\mathbf{H}(x)^{\text{b}}}_j)} \tag{8}$$

the other approach is analysis of the posterior residuals. Desroziers et al. (2005) noted that the residuals must be consistent with the PDF assumed for the model-data mismatch, here described by $\mathbf{R}$. If this is not the case we ~~can make an ad hoc~~ need to make a correction to $\mathbf{R}$~~to include~~. Here again we have a range of choices. If we have enough data we can fit covariance models as functions of space and time. We do not have enough data so we calculate directly the ensemble covariance of the residuals ~~calculated~~ as

$$\mathbf{R}_{i,j}^{\text{sample}} = \overline{(\mathbf{H}\mathbf{x}_i^{\text{a}} - \mathbf{y}_i)(\mathbf{H}\mathbf{x}_j - \mathbf{y}_j)} \tag{9}$$

where the overbar denotes an average over the ensemble of models and their respective analyses and the indices i and j refer to observations. Descriptively $\mathbf{R}^{\text{sample}}$ will be positive if, on average, models make errors of the same sign for observations i and j. Note that if the ensemble of models is smaller than the number of observations (usually the case) then both $\mathbf{R}^{\text{sample}}$ ~~is~~ and $\mathbf{R}^{\text{prior}}$ are singular. This is one reason why we ~~sum~~ add $\mathbf{R}$ ~~and $\mathbf{R}^{\text{sample}}$~~ to either, the other being that the ~~residuals do~~ model uncertainty does not capture all the data uncertainty. We note in advance an objection to using $\mathbf{R}^{\text{sample}}$ that, by using the residuals, we are double counting information in any subsequent inversion. This is partly true although firstly we only use it to correct the spread not the location of the related PDFs and that the same objection holds for any use of posterior diagnostics. The first-guess and residual covariances from Eq. 9 and Eq. 8 show somewhat similar structure, with the largest values for a few terrestrially-influenced stations such as Baltic Sea, Hungary and taiane Peninsula, Korea. As expected the variances in Eq. 9 are smaller than those in Eq. 8 reflecting the convergence of simulations towards the observations.

The weights for the case considering covariance of ~~residuals~~ first guesses is shown as the bottom row in Figure 1 and the impact on regional estimates is shown as the right bar in Figure 2. The ranking is similar to the other cases, especially for the preferred models. The main effect of including the residual covariance is to reduce the penalty for the least preferred models. Given the small changes among the preferred models it is no surprise that there is little change in the regional estimates or total uncertainties. One reason for the largest impact falling on the least preferred models is that the residual covariance is dominated by the largest residuals which come from the least preferred models.

5 Model Comparison and Cross-Validation

In Section 3 we applied the theory to the simplest possible case of models with identical dimensionality and uncertainties; they differed only in their Green's Function. The theory is more general than this. We noted in Section 2.1 that model performance is determined by the normalised prediction error and the volume of the data space occupied by the prior model. Neither of these depends directly on the dimensionality of the prior model. We can compare a model with two highly uncertain parameters against another with four more certain parameters. This extends the BIC which considers only the number of parameters.

The case is quite common in biogeochemistry in which we often compare simple models with empirical and highly uncertain parameters with complex, physically-based models whose parameters can be linked to field experiments.

A special case occurs when we compare the prior and posterior models. This is usually done by holding back a subset of the data and testing the improvement in the fit to that data (e.g. Peylin et al., 2016). The approach is frequently called cross-validation. The JIC provides a good basis for comparison of the prior and posterior models. Most importantly it accounts for the different volumes in the data space occupied by the prior and posterior models. Posterior models (informed by the previous assimilation) always occupy less volume in the space of the cross-validation data than their unconstrained or free-running prior model. Thus a good fit to the cross-validation data is less likely to be a chance event.

It is also possible to weight model estimates by their ability to fit cross-validation data. The steps are as follows:

1. Divide data into assimilation and validation data;
2. Carry out an ensemble of assimilations using each model and the assimilation data;
3. Calculate the JIC using the *posterior* estimates from step two and the validation data;
4. Calculate ensemble statistics from the posterior estimates from step two and the JIC from step three.

Note that the prior means and covariances in Equation 4 for step three are the posterior means and covariances from step two. Thus, while in Section 3.1 we varied only the model $\mathbf{H}$ here we also vary X^{b} and $\mathbf{B}$. Variations in $\mathbf{B}$ or, more generally, variations in the projection of prior uncertainty into observation space are not usually treated in cross-validation studies (e.g. Pickett-Heaps et al., 2011).

For our example we parallel the test of Stephens et al. (2007). They held back data from airborne profiles and rated models according to their ability to fit seasonal changes in vertical gradients. We cannot use the same measure in our annual mean experiment but we do use the nine points from the airborne profiles above Cape Grim Tasmania or Colorado USA.

We can calculate the JIC using these nine measurements and the prior and posterior models. The comparison of the unnormalised JIC for these cases shows whether the fit to the data held back from the inversion has improved. One would hope so but Peylin et al. (2016) showed that this is not always the case. In our case the unnormalised JIC improves by several orders of magnitude due both to a reduction in the residuals and a narrowing of the PDF. Figure 3 shows the comparison of the normalised JIC for the prior (top) and posterior (bottom) models. The prior case shows little variation around the equally-weighted value of $\frac{1}{17}$ while this variation is considerably increased for the posterior case. Figure 4 shows the ensemble statistics for three inversion cases. The left bar is the equally weighted case for the entire network (the left bar from Figure 2), the middle bar shows the equally weighted case for the inversion with the nine cross-validation stations removed while the right bar shows the same inversion but weighted according to the JIC from the cross-validation data. Averaged across all regions the impact of changing network and changing weighting are comparable although the largest changes are in North and South America following from the change of network. This was also observed by Pickett-Heaps et al. (2011).

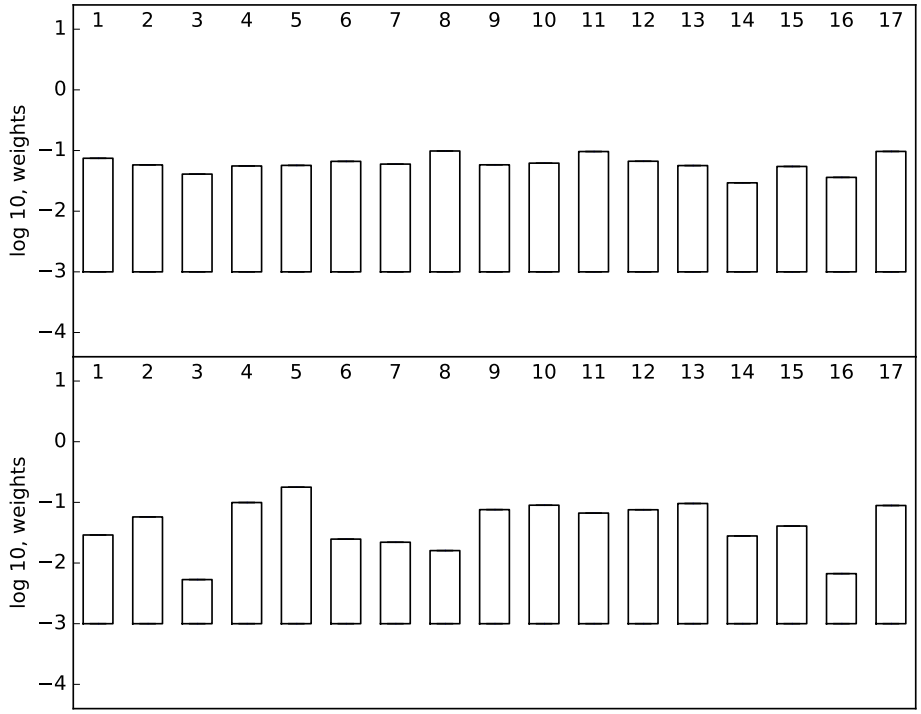

Figure 3. $\log_{10}$ of ~~w_i~~ $P(\mathbf{H}_i|\mathbf{y})$ for the prior (top) posterior (bottom) with the JIC calculated using nine airborne measurements over Cape Grim and Colorado.

6 Computational Aspects

The hardest part of the calculation of the JIC is calculating the matrix $\mathbf{HBH}^T + \mathbf{R}$. There are several possible routes depending on the size of the problem and the available machinery. In problems with few parameters it may be possible to calculate and store $\mathbf{H}$ directly. Recall that $\mathbf{H} = \nabla_{\mathbf{x}}\mathbf{y}$. We can calculate $\mathbf{H}$ either as the tangent linear of M (Griewank, 2000) or via finite difference calculations in which a parameter is perturbed. Once we calculate $\mathbf{H}$ we can generate the eigen-values of $\mathbf{HBH}^T + \mathbf{R}$ from the singular values of $\mathbf{H}$.

If the problem is too large or the generation of the Jacobian too costly we need to generate an approximation ~~to~~ of the determinant of $\mathbf{HBH}^T + \mathbf{R}$~~by calculating eigen-vectors from~~. A common approach is to calculate the leading eigenvalues of (the symmetric matrix $\mathbf{HBH}^T$ through a so-called matrix-free approach. Rather than an explicit representation of the matrix, matrix-free approaches require the capability to evaluate the product of the matrix in question with any given vector. The prime example of a matrix free approach was published by Lanczos (1950). In our case the application of a matrix-free approach requires the tangent linear ~~model for H~~of $H($ for multiplication with $\mathbf{H}(\mathbf{x})$ and the adjoint model for multiplication with $\mathbf{H}^{\mathrm{t}}(\mathbf{x})$. This is similar to calculations performed in the ~~solution of~~ conjugate gradient algorithm for the assimilation problem itself (Fisher, 1998). The second term in Equation 4 is the Bayesian least squares cost function evaluated at the minimum so, provided we want to calculate X^{a} and not just $P(\mathbf{H}_i)$ we already have this value.

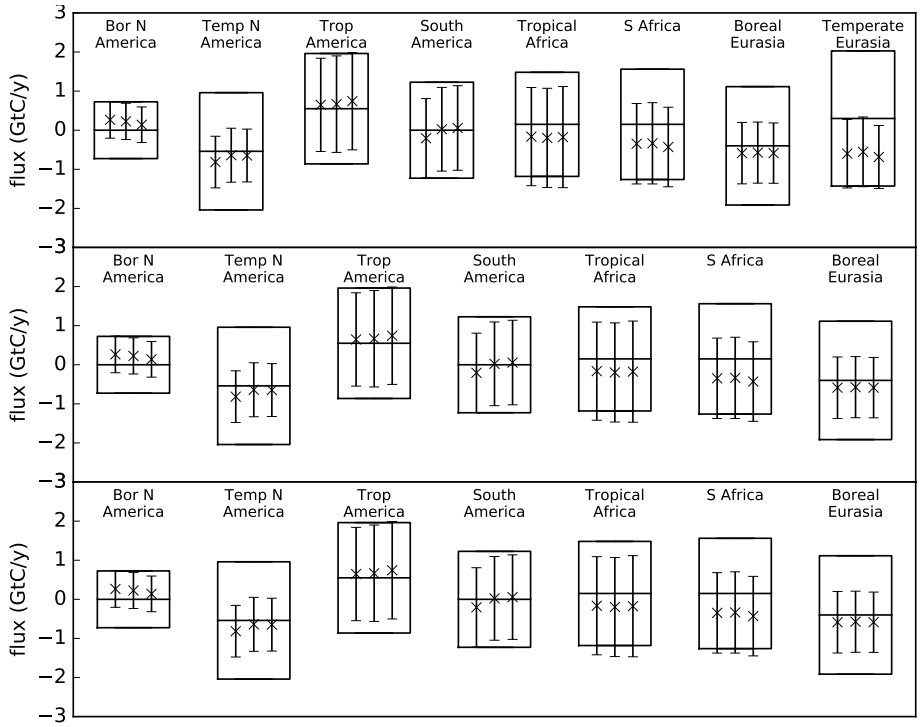

Figure 4. Prior and posterior uncertainties for regional fluxes from the TRANSCOM intercomparison following Gurney et al. (2002). The centre line of each box shows the prior estimate of the mean while the box limits show the $\pm 1\sigma$ uncertainties. The three bars show the mean (marked with "x") and $\pm 1\sigma$ uncertainty denoted by the length of the bar. The uncertainty is that of the ensemble including both the uncertainty for each model and the dispersion among model means. The left bar shows the equally weighted case for the full network, the middle bar the equally weighted case with the cross-validation stations removed and the right bar the JIC-weighted case for the cross-validation data.

7 Discussion and Future Work

The method we have outlined points out one way of incorporating measures of model quality into ensemble estimates. The TRANSCOM case points out its main limitation, a strong dependence on the underlying PDFs. The same limitation holds for other calculations with the underlying PDFs, especially measures of information content or posterior uncertainty. ~~In cases like~~ Thus the largest effort needed to improve our calculation is the same as that for many other aspects of assimilation, namely the assessment of the independent information available from large sets of observations, accounting for systematic errors in observation operators. This problem is particularly difficult in biogeochemical assimilation. the normal application is of a single assimilation carried out over the longest possible period. This is desirable both because there is usually little data available in any period (encouraging maximising the assimilation window) and many of the processes we seek to elucidate are slow so that long windows are desirable to reveal them. This means that it is hard to separate systematic errors arising from the prior, the data itself or the observation operator.

Some assimilation problems are less subject to this weakness. In numerical weather prediction~~where~~, for example, we have repeat assimilations~~we can easily~~. Thus we can test that the underlying PDFs are consistent with their realisations~~but we~~. We also have more direct tests of the quality of the assimilation via forecast skill. The above argument suggests a strong need for ensemble approaches in biogeochemical assimilation.

A more immediate application than properly weighting an ensemble of models may ~~well be in development where we need to test whether the extra complexity of one version~~ be in model development. Here a common question is of complexity over simplicity. If, as is argued throughout this series, assimilation is a good guide to parameter choice and even structure in models we need some way to tell whether adding extra processes, with their concomitant uncertainties, is worth the effort. This is a standard problem in statistical inference. the Bayesian formulation outlined here shifts the comparison of two models from complexity to the volume of data space available to them, allowing both complexity and uncertainty to play a role. this offers a promising basis for comparing different versions of a model.

8 Conclusions

We have developed a simple application of hierarchical data assimilation to incorporate choice among an ensemble of models. We have demonstrated it for a computationally simple case, the annual mean version of the TRANSCOM intercomparison. The method provides unrealistically strong discrimination among models, mainly due to incorrect assumptions about underlying PDFs. We have also successfully applied the technique to the cross-validation of the TRANSCOM inversions by holding back airborne data over Tasmania and Colorado. The method, when coupled with more sophisticated diagnostics of model-data mismatch should prove a useful extension to traditional biogeochemical data assimilation.

Code and Data Availability

The code and data files to run the TRANSCOM example and generate the figures in the paper can be found at https://figshare.com/articles/Code_needed_to_run_the_transcom_ensemble_weighted_probability_case_for_Data_Assimilation_using_an_Ensemble_of_Models_A_hierarchical_approach_Geoscience_Model_Development_Discussions_2016_w_draft_item/4210212

9 ~~Appendix A: Finding the Weights~~

~~We proceed via the multiplication of PDFs described in (Rayner et al., 2016, Section 4). We start with a uniform prior distribution for the choice of our N models $K(\mathbf{H}_i) = \frac{1}{N}$ and Gaussian PDFs for prior estimates of parameters and for data. Our problem consists of finding the marginal probability $P(\mathbf{H} = \mathbf{H}_i)$.~~

~~Using Eq. 1.93 from Tarantola (2005) we have~~

$$p(\mathbf{H}_i, \mathbf{x}) = \frac{K(\mathbf{H}_i)G(\mathbf{x}, \mathbf{x}^{\mathrm{b}}, \mathbf{P}) \cdot G(\mathbf{H}_i\mathbf{x}, \mathbf{y}, \mathbf{R})}{\sum_i \int K(\mathbf{H}_i)G(\mathbf{x}, \mathbf{x}^{\mathrm{b}}, \mathbf{P})G(\mathbf{H}_i\mathbf{x}, \mathbf{y}, \mathbf{R})dx}$$

We wish to find

$$P(\mathbf{H}=\mathbf{H}_i)=\int \sigma(\mathbf{H}_i,\mathbf{x})dx$$

and also

$$P(\mathbf{x})=\sum_i p(\mathbf{x},\mathbf{H}_i)$$

8.0.1 Model Probability

the denominator is a normalization so if we worry only about relative likelihoods we need only the integral of the numerator. The multivariate Gaussian can be expanded as:

$$G(\mathbf{x}\mu,\mathbf{C})=(2\pi)^{-n/2}|\mathbf{C}|^{-1/2}\exp\left[-\frac{1}{2}(\mathbf{x}-\mu)^T\cdot\mathbf{C}^{-1}\cdot(\mathbf{x}-\mu)\right]$$

where n is the dimension of μ.

Substituting Eq. ?? into Eq. ?? gives

$$p(\mathbf{H}_i,\mathbf{x})\propto|\mathbf{P}|^{-1/2}|\mathbf{R}|^{-1/2}\exp\left[-\frac{1}{2}(\mathbf{x}-\mathbf{x}^{\mathrm{b}})^T\cdot\mathbf{P}^{-1}\cdot(\mathbf{x}-\mathbf{x}^{\mathrm{b}})\right]\exp\left[-\frac{1}{2}(\mathbf{H}\mathbf{x}-\mathbf{y})^T\cdot\mathbf{R}^{-1}\cdot(\mathbf{H}\mathbf{x}-\mathbf{y})\right]$$

To simplify this expression we note that determinant is distributive over multiplication and also that multiplying exponentials is achieved by adding exponents. Some linear algebra, completing the square and the use of the special form of the matrix inversion lemma

$$(\mathbf{P}^{-1}+\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H})^{-1}=\mathbf{P}-\mathbf{P}\mathbf{H}^T(\mathbf{H}\mathbf{P}\mathbf{H}^T+\mathbf{R})^{-1}\mathbf{H}\mathbf{P}$$

we can reduce Eq. ?? to the form

$$p(\mathbf{H}_i,\mathbf{x})\propto|\mathbf{P}|^{-1/2}|\mathbf{R}|^{-1/2}\exp\left[-\frac{1}{2}(\mathbf{x}-\mu)^T\cdot\mathbf{A}^{-1}\cdot(\mathbf{x}-\mu)\right]\times\exp\left[-\frac{1}{2}(\mathbf{y}-\mathbf{H}_i\mathbf{x}^{\mathrm{b}})^T\cdot(\mathbf{R}+\mathbf{H}_i\,mathbf{P}\mathbf{H}_i^T)^{-1}\cdot(\mathbf{y}-\mathbf{H}_i\mathbf{x}^{\mathrm{b}})\right]$$

where

$$\mu=\mathbf{x}^{\mathrm{b}}+\mathbf{P}\mathbf{H}_i^T\cdot(\mathbf{R}+\mathbf{H}_i\mathbf{P}\mathbf{H}_i^T)^{-1}\cdot(\mathbf{y}-\mathbf{H}_i\mathbf{x}^{\mathrm{b}})$$

and

$$\mathbf{A}^{-1}=\mathbf{P}^{-1}+\mathbf{H}_i^T\mathbf{R}^{-1}\mathbf{H}_i$$

Note that Eqs. ?? and ?? are the standard expressions for the posterior mean and variance of $mathbfx$. Now substituting in Eq. ?? and using the fact that

$$\int dx\exp\left[-\frac{1}{2}(\mathbf{x}-\mu)^T\cdot\mathbf{A}^{-1}\cdot(\mathbf{x}-\mu)\right]=|\mathbf{A}|^{1/2}(2\pi)^{n/2}$$

~~we have~~

$$p(\mathbf{H}_i) \propto \left|\mathbf{A}\mathbf{P}^{-1}\right| \left|\mathbf{R}^{-1}\right| \exp\left[-\frac{1}{2}(\mathbf{y}-\mathbf{H}_i\mathbf{x}^{\mathrm{b}})^T \cdot (\mathbf{R}+\mathbf{H}_i\mathbf{P}\mathbf{H}^T)^{-1} \cdot (\mathbf{y}-\mathbf{H}_i\mathbf{x}^{\mathrm{b}})\right]$$

~~We also have the condition that~~

$$\sum_i p(\mathbf{H}_i) = 1.$$

~~Finally we can simplify $\left|\mathbf{A}\mathbf{P}^{-1}\right| \left|\mathbf{R}^{-1}\right|$ using Eq. **??** and Sylvester's Determinant Theorem which states that for any matrices $\mathbf{U}$ and $\mathbf{V}$~~

$$|\mathbf{I}+\mathbf{U}\mathbf{V}| = |\mathbf{I}+\mathbf{V}\mathbf{U}|$$

~~Substituting and simplifying yields~~

$$p(\mathbf{H}_i) \propto \left|\mathbf{R}+\mathbf{H}_i\mathbf{P}\mathbf{H}_i^{\mathrm{T}}\right|^{-1/2} \exp\left[-\frac{1}{2}(\mathbf{y}-\mathbf{H}_i\mathbf{x}^{\mathrm{b}})^T \cdot (\mathbf{R}+\mathbf{H}_i\mathbf{P}\mathbf{H}_i^T)^{-1} \cdot (\mathbf{y}-\mathbf{H}_i\mathbf{x}^{\mathrm{b}})\right]$$

8.1

Acknowledgements. this work was partly supported by an Australian Professorial Fellowship (DP1096309). We acknowledge the support from the International Space Science Institute (ISSI). This publication is an outcome of the ISSI's Working Group on "Carbon Cycle Data Assimilation: How to Consistently Assimilate Multiple Data Streams".

References

Akaike, H.: A new look at the statistical model identification, IEEE transactions on automatic control, 19, 716–723, 1974.

Baker, D. F., Law, R. M., Gurney, K. R., Rayner, P., Peylin, P., Denning, A. S., Bousquet, P., Bruhwiler, L., Chen, Y.-H., Ciais, P., Fung, I. Y., Heimann, M., John, J., Maki, T., Maksyutov, S., Masarie, K., Prather, M., Pak, B., Taguchi, S., and Zhu, Z.: TransCom 3 inversion intercomparison: Impact of transport model errors on the interannual variability of regional CO_2 fluxes, 1988–2003, Global Biogeochem. Cycles, 20, GB1002, doi:10.1029/2004GB002439, 2006.

Bodman, R. W., Rayner, P. J., and Karoly, D. J.: Uncertainty in temperature projections reduced using carbon cycle and climate observations, Nature Climate Change, doi:10.1038/NCLIMATE1903, 2013.

Cressie, N., Calder, C. A., Clark, J. S., Hoef, J. M. V., and Wikle, C. K.: Accounting for uncertainty in ecological analysis: the strengths and limitations of hierarchical statistical modeling, Ecological Applications, 19, 553–570, doi:10.1890/07-0744.1, http://dx.doi.org/10.1890/07-0744.1, 2009.

Desroziers, G., Berre, L., Chapnik, B., and Poli, P.: Diagnosis of observation, background and analysis-error statistics in observation space, Quarterly Journal of the Royal Meteorological Society, 131, 3385–3396, doi:10.1256/qj.05.108, http://dx.doi.org/10.1256/qj.05.108, 2005.

Fisher, M.: Minimization algorithms for variational data assimilation, in: Proc. ECMWF Seminar on Recent Developments in Numerical Methods for Atmospheric Modelling, pp. 364–385, Reading, 1998.

Friedlingstein, P., Cox, P., Betts, R., Bopp, L., von Bloh, W., Brovkin, V., Doney, S., Eby, M., Fung, I., Govindasamy, B., John, J., Jones, C., Joos, F., Kato, T., Kawamiya, M., Knorr, W., Lindsay, K., Matthews, H. D., Raddatz, T., Rayner, P., Reick, C., Roeckner, E., Schnitzler, K.-G., Schnur, R., Strassmann, K., Thompson, S., J.Weaver, A., Yoshikawa, C., and Zeng, N.: Climate -carbon cycle feedback analysis, results from the C4MIP model intercomparison, J. Clim., 19, 3737–3753, doi:10.1175/JCLI3800.1, 2006.

Ganesan, A. L., Rigby, M., Zammit-Mangion, A., Manning, A. J., Prinn, R. G., Fraser, P. J., Harth, C. M., Kim, K.-R., Krummel, P. B., Li, S., Mühle, J., O'Doherty, S. J., Park, S., Salameh, P. K., Steele, L. P., and Weiss, R. F.: Characterization of uncertainties in atmospheric trace gas inversions using hierarchical Bayesian methods, Atmospheric Chemistry and Physics, 14, 3855–3864, doi:10.5194/acp-14-3855-2014, http://www.atmos-chem-phys.net/14/3855/2014/, 2014.

Griewank, A.: Evaluating Derivatives: Principles and Techniques of Automatic Differentiation, SIAM, Philadelphia, Pa., 2000.

Gurney, K. R., Law, R. M., Denning, A. S., Rayner, P. J., Baker, D., Bousquet, P., Bruhwiler, L., Chen, Y.-H., Ciais, P., Fan, S., Fung, I. Y., Gloor, M., Heimann, M., Higuchi, K., John, J., Maki, T., Maksyutov, S., Masarie, K., Peylin, P., Prather, M., Pak, B. C., Randerson, J., Sarmiento, J., Taguchi, S., Takahashi, T., and Yuen, C.-W.: Towards robust regional estimates of CO_2 sources and sinks using atmospheric transport models, Nature, 415, 626–630, 2002.

Gurney, K. R., Law, R. M., Denning, A. S., Rayner, P. J., Baker, D., Bousquet, P., Bruhwiler, L., Chen, Y.-H., Ciais, P., Fan, S., Fung, I. Y., Gloor, M., Heimann, M., Higuchi, K., John, J., Kowalczyk, E., Maki, T., Maksyutov, S., Peylin, P., Prather, M., Pak, B. C., Sarmiento, J., Taguchi, S., Takahashi, T., and Yuen, C.-W.: TransCom 3 CO_2 inversion intercomparison: 1. Annual mean control results and sensitivity to transport and prior flux information, Tellus, 55B, 555–579, doi:10.1034/j.1600-0560.2003.00049.x, 2003.

Gurney, K. R., Law, R. M., Denning, A. S., Rayner, P. J., Pak, B. C., Baker, D., Bousquet, P., Bruhwiler, L., Chen, Y.-H., Ciais, P., Fung, I. Y., Heimann, M., John, J., Maki, T., Maksyutov, S., Peylin, P., Prather, M., and Taguchi, S.: Transcom 3 inversion intercomparison: Model mean results for the estimation of seasonal carbon sources and sinks, Global Biogeochem. Cycles, 18, GB1010, doi:10.1029/2003GB002111, 2004.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T.: Bayesian Model Averaging: A Tutorial, Statistical Science, 14, 382–401, http://www.jstor.org/stable/2676803, 1999.

Jaynes, E. and Bretthorst, G.: Probability Theory: The Logic of Science, Cambridge University Press, http://books.google.com.au/books?id=tTN4HuUNXjgC, 2003.

Kass, R. E. and Raftery, A. E.: Bayes factors, Journal of the american statistical association, 90, 773–795, 1995.

Kuppel, S., Chevallier, F., and Peylin, P.: Quantifying the model structural error in carbon cycle data assimilation systems, Geoscientific Model Development, 6, 45–55, doi:10.5194/gmd-6-45-2013, http://www.geosci-model-dev.net/6/45/2013/, 2013.

Lanczos, C.: An iteration method for the solution of the eigenvalue problem of linear differential and integral operators, J. Res. Natl. Bur. Stand. B, 45, 255–282, doi:10.6028/jres.045.026, 1950.

Lauvaux, T., Pannekoucke, O., Sarrat, C., Chevallier, F., Ciais, P., Noilhan, J., and Rayner, P. J.: Structure of the transport uncertainty in mesoscale inversions of CO_2 sources and sinks using ensemble model simulations, Biogeosciences, 6, 1089–1102, 2009.

MacKay, D. J. C.: Information Theory, Inference, and Learning Algorithms, Cambridge University Press, http://www.cambridge.org/0521642981, available from `http://www.inference.phy.cam.ac.uk/mackay/itila/`, 2003.

Michalak, A. M., Hirsch, A., Bruhwiler, L., Gurney, K. R., Peters, W., and Tans, P. P.: Maximum likelihood estimation of covariance parameters for Bayesian atmospheric trace gas surface flux inversions, J. Geophys. Res., 110, D24 107, doi:10.1029/2005JD005970, 2005.

Murphy, J. M., Booth, B. B., Collins, M., Harris, G. R., Sexton, D. M., and Webb, M. J.: A methodology for probabilistic predictions of regional climate change from perturbed physics ensembles, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 365, 1993–2028, 2007.

Peylin, P., Bacour, C., MacBean, N., Leonard, S., Rayner, P., Kuppel, S., Koffi, E., Kane, A., Maignan, F., Chevallier, F., Ciais, P., and Prunet, P.: A new stepwise carbon cycle data assimilation system using multiple data streams to constrain the simulated land surface carbon cycle, Geoscientific Model Development, 9, 3321–3346, doi:10.5194/gmd-9-3321-2016, http://www.geosci-model-dev.net/9/3321/2016/, 2016.

Pickett-Heaps, C. A., Rayner, P. J., Law, R. M., Bousquet, P., Peylin, P., Patra, P., Maksyutov, S., Marshall, J., Rödenbeck, C., Ciais, P., Langenfelds, R., Tans, P., Steele, P., and Francey, R.: Atmospheric CO_2 Inversion Cross-Validation Using Vertical Profile Measurements: Analysis of Four Independent Inversion Models, J. Geophys. Res., 116, D12 305, doi:10.1029/2010JD014887, 2011.

Raftery, A. E., Gneiting, T., Balabdaoui, F., and Polakowski, M.: Using Bayesian Model Averaging to Calibrate Forecast Ensembles, Monthly Weather Review, 133, 1155–1174, doi:10.1175/MWR2906.1, http://dx.doi.org/10.1175/MWR2906.1, 2005.

Rayner, P., Michalak, A. M., and Chevallier, F.: Fundamentals of Data Assimilation, Geoscientific Model Development Discussions, 2016, 1–21, doi:10.5194/gmd-2016-148, http://www.geosci-model-dev-discuss.net/gmd-2016-148/, 2016.

Rayner, P. J.: Optimizing CO_2 observing networks in the presence of model error: results from TransCom 3, Atmos. Chem. Phys., 4, 413–421, 2004.

Rayner, P. J., Koffi, E., Scholze, M., Kaminski, T., and Dufresne, J.-L.: Constraining predictions of the carbon cycle using data, Phil. Trans. Roy. Soc. A, 369, 1955–1966, doi:10.1098/rsta.2010.0378, 2011.

Scholze, M., Kaminski, T., Rayner, P., Knorr, W., and Geiring, R.: Propagating uncertainty through prognostic CCDAS simulations, J. Geophys. res., 112, d17 305, doi:10.1029/2007JD008642, 2007.

Schwarz, G.: Estimating the Dimension of a Model, Ann. Statist., 6, 461–464, doi:10.1214/aos/1176344136, http://dx.doi.org/10.1214/aos/1176344136, 1978.

Stephens, B. B., Gurney, K. R., Tans, P. P., Sweeney, C., Peters, W., Bruhwiler, L., Ciais, P., Ramonet, M., Bousquet, P., Nakazawa, T., Aoki, S., Machida, T., Inoue, G., Vinnichenko, N., Lloyd, J., Jordan, A., Heimann, M., Shibistova, O., Langenfelds, R. L., Steele, L. P., Francey, R. J., and Denning, A. S.: Weak Northern and Strong Tropical Land Carbon Uptake from Vertical Profiles of Atmospheric CO_2, Science, 316, 1732–1735, doi: 10.1126/science.1137004, 2007.

Stordal, A. S., Karlsen, H. A., Nævdal, G., Skaug, H. J., and Vallès, B.: Bridging the ensemble Kalman filter and particle filters: the adaptive Gaussian mixture filter, Computational Geosciences, 15, 293–305, 2011.

Takahashi, T., Wanninkhof, R. H., Feely, R. A., Weiss, R. F., Chipman, D. W., Bates, N., Olafsson, J., Sabine, C., and Sutherland, S. C.: Net sea-air CO_2 flux over the global oceans: An improved estimate based on the sea-air pCO_2 difference, in: Extended abstracts of the 2nd International CO_2 in the Oceans Symposium, edited by Nojiri, Y., pp. 9–15, National Institute for Environmental Studies, 1999.

Tarantola, A.: Inverse Problem Theory: Methods for Data Fitting and Parameter Estimation, Elsevier, Amsterdam, 1987.

Tarantola, A.: Inverse Problem Theory and Methods for Model Parameter Estimation, SIAM, (ISBN 0-89871-572-5), 2005.

Taylor, K. E., Stouffer, R. J., and Meehl, G. A.: An overview of CMIP5 and the experiment design, Bulletin of the American Meteorological Society, 93, 485, 2012.