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).

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^4MIP (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, 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 CO2 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 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 review the necessary machinery. 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

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 target variables are implicitly dependent on the observation operator. Where an ensemble of observation operators is available we 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 observation operators and unknowns using an expression known variously as the chain rule of probabilities or the law of total probabilities

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

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 observation operators provided they are approximately linear over enough of the support for the joint distribution of $\mathbf{x}$ and $\mathbf{y}$. The qualitative ranking of models is unlikely to be sensitive to weak nonlinearities since, as we shall see, the discrimination among models is strong.

We follow the notation of Rayner et al. (2016). Take a collection of linear observation operators with Jacobians $\mathbf{H}_1 \dots \mathbf{H}_N$, with prior probability for the unknowns given by $G(\mathbf{x}|\mathbf{x}^{\mathrm{b}},\mathbf{B})$ and prior probability for the data given by $G(\mathbf{y}|\mathbf{y}^{\mathrm{o}},\mathbf{R})$ where $G(\mathbf{x}|\mu,\mathbf{C})$ represents the Gaussian distribution of the variable $\mathbf{x}$ given mean μ and uncertainty covariance $\mathbf{C}$.

For each $\mathbf{H}_i$ our problem is the linear Gaussian inversion described in (Rayner et al., 2016, Section 6.4). Most importantly for us 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

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

[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.

where $\mathbf{x}_i^{\mathrm{a}}$ 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 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)

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

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 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 $P(\mathbf{H}_i)$ (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 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. 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 in Section 4.

2.2 Relationship with Other Criteria

$P(\mathbf{H}_i)$ is related to several other measures of model quality. For convenient comparison we define

$$\mathrm{L} = -2\log\left(\frac{P(\mathbf{H}_i|\mathbf{y})}{K}\right) = \log\left|\mathbf{H}_i\mathbf{B}\mathbf{H}_i^T+\mathbf{R}\right| + \chi^2 \quad (5)$$

The change of sign means smaller values of L correspond to more likely models.

L is related to 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). Both these criteria penalise models for adding parameters. Neither take account of different prior uncertainties among parameters or different sensitivities of the observations to these parameters.

3 The TRANSCOM Example

The **TRANSCOM** III intercomparison (Gurney et al., 2002, 2004; Baker et al., 2006) was designed to investigate the impact of uncertainty in atmospheric transport models on the determination of CO2 sources and sinks. The target variables were the mean CO2 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 CO2 concentration from 77 stations and the same data was used in every inversion. 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 the posterior fluxes such as the mean, the mean uncertainty (averaging all the posterior uncertainties) and finally the "between model" spread, calculated as the covariance among the posterior fluxes for each model. In all these calculations we weighted every model equally. What happens if we apply the methods described in Section 2 to calculate pooled estimates?

Figure 1 shows a slightly modified L 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 $\mathbf{P}$ 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 L) 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.

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 assumption that we must choose one (see Jaynes and Bretthorst, 2003, P136 for a discussion of this point). We must, however, be careful not to over-interpret these probabilities as measures of model quality. In the first place, L, 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

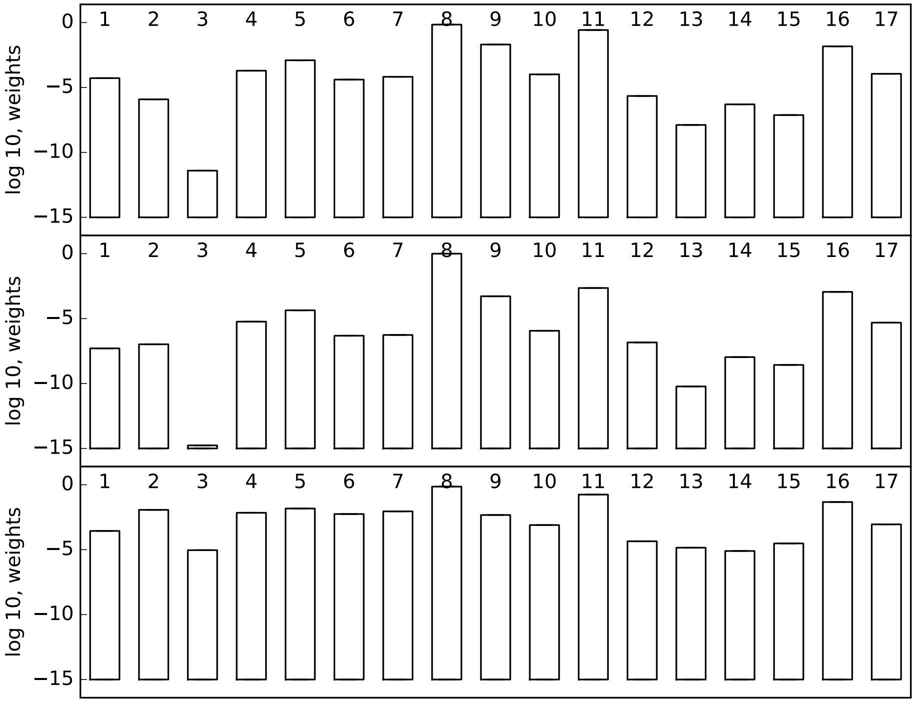

Figure 1. $\log_{10}$ of $P(\mathbf{H}_i|\mathbf{y})$ for the untuned (top), tuned (middle) and case with residuals used for $\mathbf{R}$ (bottom) transcom inversions.

exponentials to calculate probabilities. 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 normalise by considering L/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

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

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

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

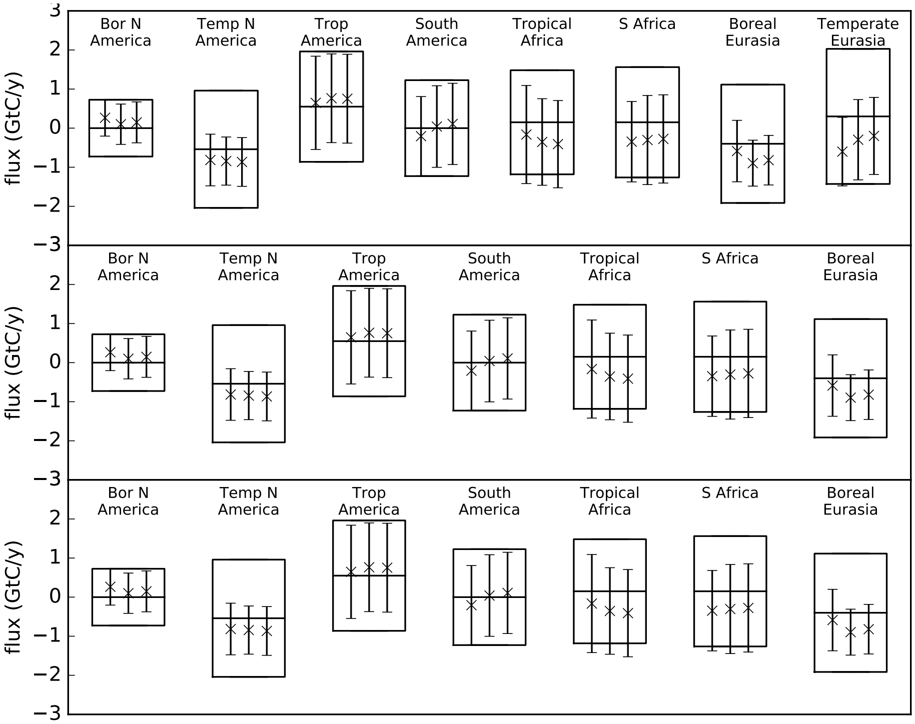

Figure 2. Prior and posterior uncertainties for regional fluxes from the TRANSCOM intercomparison following Gurney et al. (2002). The centre line of each box shows the prior estimate of the mean while the box limits show the $\pm 1\sigma$ uncertainties. The three bars show the mean (marked with "x") 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 $P(\mathbf{H}_i|\mathbf{y})$ and the right bar the case with covariance of residuals included.

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 $P(\mathbf{H}_i|\mathbf{y})$ 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 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 $P(\mathbf{H}_i|\mathbf{y})$ 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 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 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 by L 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 error in the formulation of L 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 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, 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 target variables as a measure of the model contribution to uncertainty. This was suggested by Tarantola (1987). We can write this as

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

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 need to make a correction to $\mathbf{R}$. 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 as

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

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 $\mathbf{R}^{\mathrm{sample}}$ will be positive if, on average, models make errors of the same sign for observations

[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

i and j. Note that if the ensemble of models is smaller than the number of observations (usually the case) then both $\mathbf{R}^{\text{sample}}$ and $\mathbf{R}^{\text{prior}}$ are singular. This is one reason why we add $\mathbf{R}$ to either, the other being that the 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 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. L 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 L using the *posterior* estimates from step two and the validation data;

4. Calculate ensemble statistics from the posterior estimates from step two and L 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 L using these nine measurements and the prior and posterior models. The comparison of L 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 L 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 normalised L 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 $P(\mathbf{H}_i|\mathbf{y}^{\mathrm{cv}})$ where $\mathbf{y}^{\mathrm{cv}}$ is 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).

6 Computational Aspects

The hardest part of the calculation of $P(\mathbf{H}_i|\mathbf{y})$ is calculating the matrix $\mathbf{H}_i\mathbf{B}\mathbf{H}_i^T + \mathbf{R}$. 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 $\mathbf{H}_i$ directly. Recall that $\mathbf{H}_i = \nabla_{\mathbf{x}}\mathbf{y}$. We can calculate $\mathbf{H}_i$ either as the tangent linear of H (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 $\mathbf{H}_i\mathbf{B}\mathbf{H}_i^T + \mathbf{R}$ from the singular values of $\mathbf{H}_i$.

If the problem is too large or the generation of the Jacobian too costly we need to generate an approximation of the determinant of $\mathbf{H}_i\mathbf{B}\mathbf{H}_i^T + \mathbf{R}$. A common approach is to calculate the leading eigenvalues of (the symmetric matrix $\mathbf{H}_i\mathbf{B}\mathbf{H}_i^T$ 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 of $H_i($ to generate $\mathbf{H}_i(\mathbf{x})$ and the adjoint model to generate $\mathbf{H}_i^{\mathrm{t}}(\mathbf{x})$. This is similar to calculations performed in the 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 $P(\mathbf{H}_i|\mathbf{y})$ we already have this value.

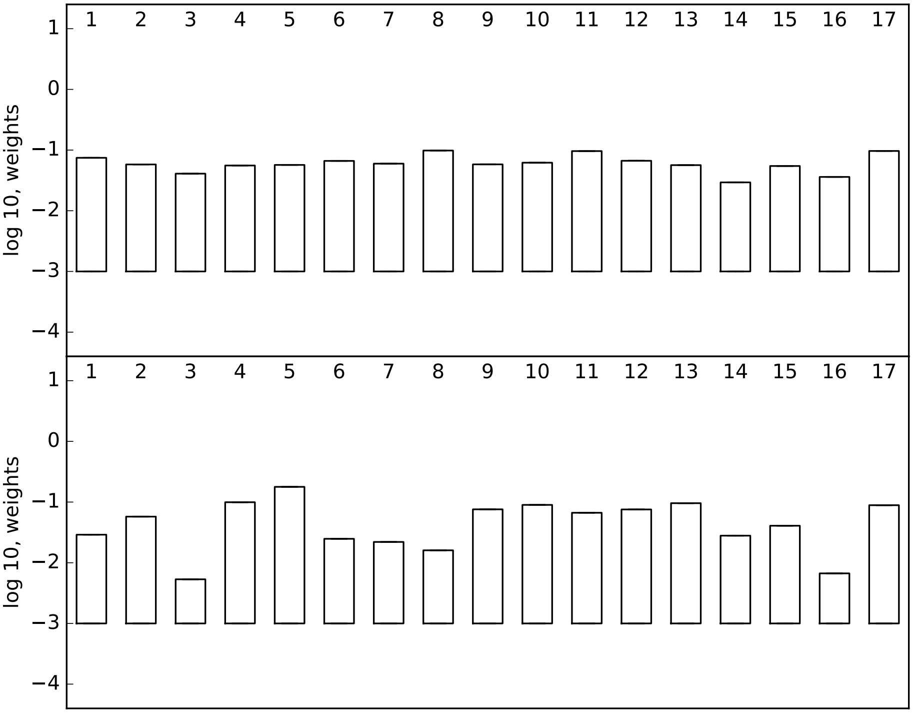

Figure 3. $\log_{10}$ of $P(\mathbf{H}_i|\mathbf{y})$ for the prior (top) posterior (bottom) with the JIC calculated using nine airborne measurements over Cape Grim and Colorado.

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. 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, for example, we have repeat assimilations. Thus we can test that the underlying PDFs are consistent with their realisations. 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.

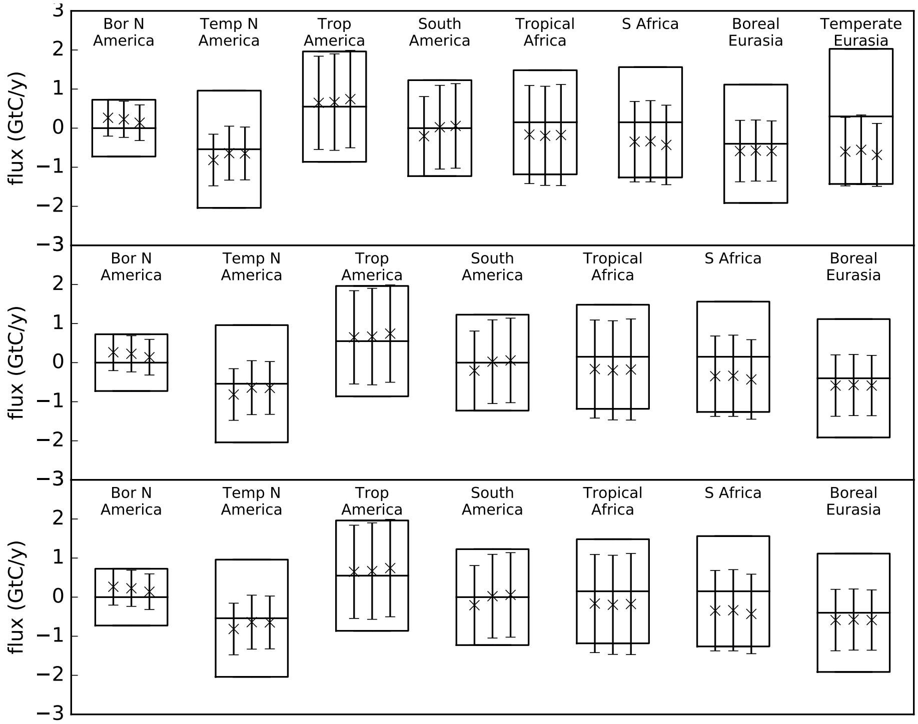

Figure 4. Prior and posterior uncertainties for regional fluxes from the TRANSCOM intercomparison following Gurney et al. (2002). The centre line of each box shows the prior estimate of the mean while the box limits show the $\pm 1\sigma$ uncertainties. The three bars show the mean (marked with "x") 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 L-weighted case for the cross-validation data.

A more immediate application than properly weighting an ensemble of models may 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.

The comparison between models and data sets is, however, incomplete. We cannot compare easily two assimilations with different amounts of data since the Model Evidence has a strong dependence on dimension.

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

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.

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.

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.