Probabilistic aggregation of individual assessment units in the U.S. Geological Survey national CO₂ sequestration assessment

Madalyn S. Blondes, John H. Schuenemeyer, Lawrence J. Drew, Peter D. Warwick

U.S. Geological Survey, Energy Resources Science Center, 12201 Sunrise Valley Dr., Reston, VA 20192, USA
Southwest Statistical Consulting, LLC, 960 Sligo St., Cortez, CO 81321, USA

Abstract

The U.S. Geological Survey is conducting a probabilistic national assessment of CO₂ storage resources in ~200 storage assessment units (SAUs), and must present results at basin, regional, and national scales. To calculate combined resources with correctly propagated uncertainty requires a probabilistic aggregation procedure in which correlations between SAUs are estimated. A comparison of single- and multiple-stage aggregation methods shows that the former better represents the varied geology within and between basins. Pairwise SAU correlation coefficients are estimated by geologists. If the matrix is not positive semi-definite, it is adjusted. Subsets of the matrix are extracted for basin or regional calculations.

© 2013 The Authors. Published by Elsevier Ltd.
Selection and/or peer-review under responsibility of GHGT

Keywords: Carbon sequestration; Natural resource assessments; Probabilistic aggregation; Monte Carlo methods

1. Introduction

Following the 2007 Energy Independence and Security Act (EISA) [1], the U.S. Geological Survey (USGS) is currently undertaking a national assessment of potential storage resources for CO₂ in sedimentary basins [2]. Geological storage of CO₂ is one potential mitigating solution to global climate change. The USGS CO₂ assessment methodology uses a Monte Carlo model to estimate the probability distribution of the mass of supercritical CO₂ that can be stored in porous sedimentary rocks [3]. Like
many large scale resource assessments, particularly for oil, gas, and minerals, the area under investigation (i.e. geologic basins of the United States) is split into smaller, more tractable units. Each of these CO₂ Storage Assessment Units (SAUs) is a formation–seal pair within a basin, and has its own probability distribution of CO₂ storage resource. Combining the probability functions for these individual units at the end of the study, with correctly propagated uncertainty, uses a method known as probabilistic aggregation [4]. Because previous oil and gas resource assessments have shown that geologic dependencies exist between assessed reservoirs, the aggregation procedure requires determining the dependencies, or correlations, between individual units [4]. Assuming that fractiles are additive (i.e. that the sum of the P₉₅ fractiles equals the P₉₅ of the sum of the storage resource distributions) is inherently an assumption of total dependence and creates an unrealistically broad uncertainty interval on the aggregated result. Fig. 1 and Table 1 show an example aggregation for three hypothetical basins A, B, and C, each with four SAUs, and for different assumptions of dependence [5]. Though the mean values are the same using all three dependency assumptions, the high estimate (P₉₅) for the independent assumption (9,736 Mt) is 56% greater than the high estimate for the dependent assumption (6,241 Mt). Clearly, understanding dependencies and aggregation are necessary for all probabilistic CO₂ storage assessments across multiple basins or countries where individual assessments are combined.

The probabilistic aggregation methodology requires two main inputs: 1) stochastic storage resource estimates for each SAU [2-3], and 2) a correlation matrix specifying dependencies between SAUs. The first section of the aggregation methodology (2.1) described below compares a single stage approach to a multiple stage approach. The second section of the methodology (2.2) describes how assessment geologists determine correlation coefficients and the mathematical aggregation procedure. This document is a summary of two recently submitted manuscripts [5-6] and uses figures directly from these papers.

![Graph showing empirical distribution](image-url)
Table 1. Summary statistics for aggregation of basins A, B, and C, as described in Fig. 1. Table reprinted from Blondes et al., accepted [5].

<table>
<thead>
<tr>
<th>Assumption</th>
<th>P05</th>
<th>P25</th>
<th>P50</th>
<th>P75</th>
<th>P95</th>
<th>mean</th>
<th>stdev</th>
<th>P95-P05</th>
<th>P75-P25</th>
</tr>
</thead>
<tbody>
<tr>
<td>Independent</td>
<td>3,488</td>
<td>4,144</td>
<td>4,656</td>
<td>5,257</td>
<td>6,241</td>
<td>4,738</td>
<td>844</td>
<td>2,753</td>
<td>1,113</td>
</tr>
<tr>
<td>User Correlation</td>
<td>2,671</td>
<td>3,684</td>
<td>4,538</td>
<td>5,577</td>
<td>7,475</td>
<td>4,738</td>
<td>1,506</td>
<td>4,804</td>
<td>1,893</td>
</tr>
<tr>
<td>Dependent</td>
<td>1,526</td>
<td>2,812</td>
<td>4,212</td>
<td>6,099</td>
<td>9,736</td>
<td>4,738</td>
<td>2,617</td>
<td>8,210</td>
<td>3,288</td>
</tr>
</tbody>
</table>

Nomenclature

SAU: Storage Assessment Unit – a geologic formation-seal pair within a sedimentary basin.

Correlation: Geologic and methodologic partial dependencies. Values are between -1 and 1.

Aggregation: A method to propagate uncertainty, using correlations, to combine individual SAUs.

Single-stage: Aggregation method using the entire or subsets of one large correlation matrix to calculate national, regional, or basin CO2 storage totals.

Multiple-stage: Aggregation method using intermediate results and multiple smaller correlation matrices to calculate national, regional, or basin CO2 storage totals.

2. USGS CO2 Aggregation methodology

2.1. Single Stage vs. multiple stage aggregation

The USGS is required to report results not only at the SAU/formation scale, but also the basin, regional, and national scales. The calculated national storage results will aggregate approximately 200 individual SAUs in approximately 30 sedimentary basins, requiring approximately 19,900 correlation coefficients. Making independent determinations of this many pairwise coefficients is difficult, particularly because of the complexity of estimating consistent coefficients between formations in geologically unrelated regions. One way to reduce the number of coefficients is to divide the probabilistic aggregation into multiple stages (Fig. 2), first aggregating at the basin, then the regional, then the national scale – a simplification that will reduce the number of correlation coefficients (Fig. 3) by more than an order of magnitude for large assessments.

To compare the final uncertainties that result from the a single stage vs. a multiple stage aggregation, it is assumed that the different permutations of stages and groupings follow the constraint that

\[ T = \prod_{i=1}^{k} n_i \]  

where \( T \) is the total number of SAUs, \( k \) is the number of stages in the aggregation, and \( n \) are the group sizes of SAUs for each stage [6]. If the group size, \( n \), is the same for each stage, equation 1 simplifies to:

\[ T = n^k. \]
The difference between single and multiple stage aggregation procedures on the aggregated uncertainty is shown in Figs. 2-3 for an example with $T = 16$ SAUs. For this example, the group size is held constant (Equation 2), and all correlation coefficients are assumed to be constant. The validity of the constant correlation coefficient assumption is addressed below. Fig. 2 shows three ways that 16 identical inputs with equal group sizes can be arranged: in one aggregation stage using a group of 16, in two aggregation stages using groups of four, or in four aggregation stages using groups of two. The corresponding correlation matrices are shown in Fig. 3. Note that from the one-stage to the four-stage example, the correlation coefficients necessary to define by expert elicitation have decreased from 120 to 15. Blondes [6] shows that as the number of stages, $k$, increases, the aggregated uncertainty narrows. This is true for all positive correlation coefficients and becomes more pronounced for larger assessment sizes, $T$, and more equal grouping sizes, $n_i$ [6]. By splitting the aggregation into multiple stages to reduce the number of independently defined correlation coefficients, the total degrees of freedom for the system have been decreased, thus artificially decreasing the final uncertainty [6].

The assumption of a common correlation coefficient between stages introduces bias into the aggregated standard deviation. Blondes (in review) [6] analytically calculates the second stage correlation coefficient choice necessary to obtain the same aggregated uncertainty as the single stage method. For the $T = 16$ example a second stage correlation coefficient of 0.5 must be chosen to arrive at the same standard deviation for a single stage aggregation with constant correlation coefficients of 0.2 [6].
Fig. 3. Correlation matrices for the multiple stage grouping scenarios shown in Fig. 2. A, B, C, and D are basins, and $A_i$, $B_i$, $C_i$, and $D_i$ are SAUs within them. (a) The top panel in blue represents a single stage aggregation with 120 individually determined correlation coefficients. (b) The middle panel in cyan represents a two stage aggregation with 30 individually determined correlation coefficients. (c) The bottom panel in green represents a four stage aggregation with only 15 individually determined correlation coefficients. For clarity and spacing concerns in this panel, only the bottom left entry of the second stage matrices are shown. However, all matrices in this panel are 2x2. Note that the smaller matrices in the middle and bottom panels are not subsets of the larger matrix in the top panel, but separate matrices for each new stage of aggregation. Figure reprinted from Blondes, in review [6].
Using the single value of 0.5 to describe a basin to basin correlation coefficient that represents SAU to SAU correlation coefficients of 0.2 makes little geologic sense. One would expect that distant sedimentary basins with no shared geologic history would have less correlation than formations in the same basin. The effect of reducing the degrees of freedom through the multiple stage approach has a much greater effect on the final uncertainty than the correlation coefficient chosen through expert geologic opinion. This effect is even more pronounced for larger assessments [6].

In many cases for this assessment, there are greater correlations across basins than within them, particularly if the same formation is found in multiple basins. Not only does this make it difficult to choose basin or regional scale coefficients that can accurately represent this varied geology, but the multiple stage simplification would necessitate a correlation coefficient much higher than is geologically reasonable. A multiple stage approach, therefore, is not the best solution for the USGS National Geologic CO$_2$ Sequestration Assessment. Instead, using the full correlation matrix for the national resource calculation, and subsets of this matrix for the regional and basin calculations, provides a valid propagation of uncertainty accounting for geologic similarities.

2.2 USGS National CO$_2$ Sequestration Aggregation Methodology

The pairwise coefficients of the national correlation matrix are specified by expert opinion [7] of the assessment geologists, taking into account dependencies that are a result of using similar analog formations to estimate geologic parameters such as porosity and thickness, and human factors related to using a single team for the entire assessment. The correlations are specified using an ordinal scale and are later mapped into pairwise correlations ranging from 0.1 to 0.7. High and medium ordinals are mapped to 0.7 and 0.5, respectively. A “formation” ordinal is used for the same formation in different basins and is mapped to a value of 0.4. Much of the large matrix can be populated with a priori assumptions about the value of correlations between distant SAUs, i.e. those in different basins or regions [5]. The low ordinal is mapped to different values depending on whether the SAUs are in the same basin (0.3), in different basins but in the same region (0.2), or in different regions (0.1). Regions are groups of basins with similar stratigraphy, tectonic history, or geographic location. The geologists may adjust these values for specific pairs if desired. Specification of pairwise correlations does not always yield a proper correlation matrix and in some instances it was necessary to apply a minor adjustment to the matrix by computing a proper correlation matrix to the nearest Frobenius norm [8]. See [5] for details.

A graphical representation of the correlation matrix used for the CO$_2$ assessment is shown in Fig. 4. This figure is expressly not to present correlation coefficients between specific SAUs, but is rather meant to show the correlation structure of a national assessment. Single regions are the large, light gray squares surrounding the diagonal, and single basins are the darker gray squares along the diagonal within them. The more highly correlated (darker gray) cells within the lighter region but outside individual basins represent the same formation found in different basins. For example, the large square region that takes up much of the top left portion of the correlation matrix represents the Rocky Mountain region. The smaller squares close to the diagonal are the Rocky Mountain basins (e.g. Bighorn and Powder River). The linear features in the large, light square are high correlations between formations in different basins within the Rocky Mountains (e.g., the Tensleep in both the Bighorn and the Powder River Basins). Once the entire correlation matrix has been established, the aggregation proceeds at the basin, regional, and national scale using the appropriate subset of the full matrix. The correlation matrices are used to induce rank correlation, using a Cholesky decomposition, among the empirical marginal distributions representing
individually assessed SAUs. The details of this calculation and the associated R code can be found in [5]. New empirical distributions of CO₂ storage resource are thus created for each aggregated basin and region, as well as the total for the United States.

3. Conclusions

Probabilistic aggregation, a method to correctly propagate uncertainty for the sum of resource estimates, is important and necessary for any natural resource assessment that is split into smaller units. The U.S. Geological Survey is conducting a probabilistic national assessment of CO₂ storage resource in ~200 individual SAUs that must be combined at the basin, region, and national levels. To calculate these combined storage resources with correctly propagated uncertainty requires the estimation of correlations, or dependencies, between SAUs. Pairwise correlation coefficients are estimated by assessment geologists and are assembled into a large correlation matrix. If the matrix is not positive semi-definite, a proper correlation matrix to the nearest Frobenius norm is computed. For aggregations to regions smaller than the entire U.S., subsets of the national correlation matrix are extracted for the calculation. This single-stage aggregation method was found to better represent the varied geology within and between basins than the multiple-stage aggregation method. The correlation matrices impart the geologic dependency structure to the individually assessed SAUs, and the aggregated storage resources are then calculated with correctly propagated uncertainty.

Fig. 4. Assessment Correlation Matrix. Each cell represents the pairwise correlation between two SAUs. The darkest cells along the diagonal have a value of 1.0. The light gray cells far from the diagonal have a baseline value of 0.1. The remaining red cells range from 0.2 to 0.7. See Section 2.2 for further explanation.
Acknowledgements

The authors thank the USGS National Geologic Carbon Sequestration assessment team, particularly Matthew Merrill and Sean Brennan for their valuable reviews.

References


