Diagnostics and optimization of analysis by cross-validation

Richard Ménard and Martijn Deshaies-Jacques
Air Quality Research Division – Environment and Climate Change Canada
richard.menard@canada.ca

Abstract. We examine how observations can be used to evaluate an air quality analysis by verifying against passive observations (i.e. cross-validation) that are not used to create the analysis and we compare these verifications to those made against the same set of (active) observations that were used to generate the analysis. The results show that both active and passive observations can be used to evaluate first moment metrics (e.g. bias) but only passive observations are useful to evaluate second moment metrics such as variance of observed-minus-analysis and correlation between observations and analysis. We derive a set of diagnostics based on passive observation minus-analysis residuals and we show that the true analysis error variance can be estimated, without relying on any statistical optimality assumption. This diagnostic is used to obtain near optimal analyses that are then used to evaluate the analysis error using several different methods. We compare the estimates according to the method of Hollingsworth Lonnberg. Desroziers, a diagnostic we introduce, and the perceived analysis error computed from the analysis scheme, to conclude that as long as the analysis is optimal, all estimates agrees within a certain error margin. The analysis error variance at passive observation sites is also obtained.

Theory

Assuming that observations errors are spatially uncorrelated and uncorrelated with background errors then the passive observation errors are uncorrelated with the cross-validation analysis error

Hilbert space representation of random variables

\[ (\mathbf{x}_i, \mathbf{y}_i) \sim \mathcal{E}(\mathbf{x} = \mathbf{x} + \mathbf{y}) \]

\[ \{ (\mathbf{o} - \mathbf{o}) \} \sim \mathcal{E}(\mathbf{b} - \mathbf{t}) \]

\[ \{ (\mathbf{o} - \mathbf{o}) - \mathbf{b} \} \sim \mathcal{E}(\mathbf{b} - \mathbf{t}) \]

\[ \{ (\mathbf{o} - \mathbf{o}) - \mathbf{b} \} \sim \mathcal{E}(\mathbf{b} - \mathbf{t}) \]

\[ \{ (\mathbf{o} - \mathbf{o}) - \mathbf{b} \} \sim \mathcal{E}(\mathbf{b} - \mathbf{t}) \]

\[ \{ (\mathbf{o} - \mathbf{o}) - \mathbf{b} \} \sim \mathcal{E}(\mathbf{b} - \mathbf{t}) \]

\[ \{ (\mathbf{o} - \mathbf{o}) - \mathbf{b} \} \sim \mathcal{E}(\mathbf{b} - \mathbf{t}) \]

In the case of optimal analysis

There are a number of methods to estimate the analysis error covariance using active observations

- Hollingsworth-Lönnberg [1] \[ \mathbb{E}((\mathbf{o} - \mathbf{o} - \mathbf{b}) - \mathbf{b} + \mathbf{b} + \mathbf{b} + \mathbf{b}) = \mathbf{R} - \mathbf{R} - \mathbf{R} + \mathbf{R} + \mathbf{R} + \mathbf{R} \]

- This study \[ \mathbb{E}((\mathbf{o} - \mathbf{o} - \mathbf{b}) - \mathbf{b} + \mathbf{b} + \mathbf{b} + \mathbf{b}) = \mathbf{R} - \mathbf{R} - \mathbf{R} + \mathbf{R} + \mathbf{R} + \mathbf{R} \]

- Desroziers [2] \[ \mathbb{E}((\mathbf{o} - \mathbf{o} - \mathbf{b}) - \mathbf{b} + \mathbf{b} + \mathbf{b} + \mathbf{b}) = \mathbf{R} - \mathbf{R} - \mathbf{R} + \mathbf{R} + \mathbf{R} + \mathbf{R} \]

- Analysis error covariance calculated by the analysis scheme – perceived analysis error

\[ \mathbb{H} - \mathbb{H} - \mathbb{H} - \mathbb{H} - \mathbb{H} - \mathbb{H} \]

Using passive observations

- This study \[ \mathbb{E}((\mathbf{o} - \mathbf{o} - \mathbf{b}) - \mathbf{b} + \mathbf{b} + \mathbf{b} + \mathbf{b}) = \mathbf{R} - \mathbf{R} - \mathbf{R} + \mathbf{R} + \mathbf{R} + \mathbf{R} \]

- General result in cross-validation space \[ \mathbb{R} + \mathbb{H} - \mathbb{H} \]

Cross-validation design

- Divide randomly the observations into 3 equal number sets

- Perform an analysis using 2 sets, and use the third one to validate the analysis. Do 3 permutations of these sets, so that all observations sites are used to validate the analysis

Experiment design

- Hourly analyses for PM10 and O3 for a period of 60 days (June 14 to August 12, 2014)

Step 1 - First guess experiment

- Maximum likelihood estimation of \( \sigma \) using second-order autoregressive model and error variances obtained from local Hollingsworth-Lönnberg fit [4]

- Analysis performed over the period of 60 days

- Using all observation “active”

- 3 cross-validation analyses using 2 subset out of the 3 with permutations

- Conduct a series of analyses with prescribed uniform error variances with different ratio \( \sigma^2 = \sigma^2 + \sigma^2 \)

Step 2 - Optimization

- Obtain the optimal \( \sigma^2 = \sigma^2 + \sigma^2 \) by minimizing var(\( \mathbf{o} - \mathbf{a} \)) that is minimizing the analysis error variance

- Reevaluate \( \mathcal{L} \) using maximum-likelihood using the new error variances

Step 3 - Optimal analysis experiment

- Re-conduct an optimization for \( \sigma^2 = \sigma^2 + \sigma^2 \)

- As in Step 1 perform a series of analysis

Results of first guess experiment (iter0) and optimization (iter1)

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Passive (( \mathbb{H} - \mathbb{H} - \mathbb{H} - \mathbb{H} - \mathbb{H} - \mathbb{H} ))</th>
<th>Passive (( \mathbb{R} + \mathbb{H} - \mathbb{H} ))</th>
<th>Passive (( \mathbb{R} + \mathbb{H} - \mathbb{H} ))</th>
</tr>
</thead>
<tbody>
<tr>
<td>O1: iter 0</td>
<td>124</td>
<td>101.25</td>
<td>83.80</td>
</tr>
<tr>
<td>O1: iter 1</td>
<td>45</td>
<td>101.25</td>
<td>83.80</td>
</tr>
<tr>
<td>PM0.1: iter 0</td>
<td>196</td>
<td>93.93</td>
<td>80.34</td>
</tr>
<tr>
<td>PM0.1: iter 1</td>
<td>86</td>
<td>93.93</td>
<td>80.34</td>
</tr>
</tbody>
</table>

Estimation of analysis error variance at the passive observation sites

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Passive (( \mathbb{R} + \mathbb{H} - \mathbb{H} ))</th>
<th>Passive (( \mathbb{R} + \mathbb{H} - \mathbb{H} ))</th>
</tr>
</thead>
<tbody>
<tr>
<td>O1: iter 0</td>
<td>26.83</td>
<td>32.72</td>
</tr>
<tr>
<td>O1: iter 1</td>
<td>28.95</td>
<td>28.75</td>
</tr>
<tr>
<td>PM0.1: iter 0</td>
<td>22.65</td>
<td>24.59</td>
</tr>
<tr>
<td>PM0.1: iter 1</td>
<td>27.62</td>
<td>21.38</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Active (( \mathbb{H} - \mathbb{H} ))</th>
<th>Active (( \mathbb{H} - \mathbb{H} ))</th>
<th>Active (( \mathbb{H} - \mathbb{H} ))</th>
</tr>
</thead>
<tbody>
<tr>
<td>O1: iter 0</td>
<td>22.69</td>
<td>9.66</td>
<td>6.03</td>
</tr>
<tr>
<td>O1: iter 1</td>
<td>13.32</td>
<td>13.68</td>
<td>8.94</td>
</tr>
<tr>
<td>PM0.1: iter 0</td>
<td>17.98</td>
<td>7.71</td>
<td>3.18</td>
</tr>
<tr>
<td>PM0.1: iter 1</td>
<td>17.98</td>
<td>7.71</td>
<td>3.18</td>
</tr>
</tbody>
</table>

Summary and Conclusions

- We examine how passive observations can be used to evaluate analyses, estimate the analysis error and to optimize the analysis

- Assuming that observation errors are horizontally uncorrelated gives the following central diagnostic [3]

- Which is valid whether or not the analysis is optimal.

- By minimizing the central diagnostic we minimize the analysis error variance thus proving a means to optimize the analysis

- The optimization of the ratio of observation error variance over background error variance, and re-evaluating the correlation length give near optimal analyses with chi-square values closer to one

- We have introduce a new diagnostic for analysis error that can be applied in both active and passive observation space

- We have compared different diagnostic of analysis error variance in active observation space, the Hollingsworth-Lönnberg [3], Desroziers [2], our diagnostic, and the computed analysis error provided by the analysis scheme and showed that they roughly agrees for near optimal analyses. Strong disagreement between the estimates is found when the analysis is not optimal.

- We have compared different diagnostic of analysis error variance in passive observation space, our diagnostic and the central diagnostic and showed strong agreement in the estimates when the analysis is optimal

The method introduced here is general and could be used in other geophysical applications and in particular in surface analyses

Study has been submitted to Atmospheric Special Issue: Air Quality Forecasting and Monitoring

References


