A-posteriori error and uncertainty estimates for DDDAS inference problems*

Adrian Sandu
Vishwas Hebbur Venkata Rao

Computational Science Laboratory
Computer Science Department
Virginia Tech

*AFOSR DDDAS FA9550-12-1-0293-DEF.
Dr. Frederica Darema
Project goal: develop computational framework for treating errors and uncertainty in DDDAS
An important DDDAS class of problems is *data assimilation*, which motivates this work.

[Singh, Sandu, et al 2010]

TES ozone column observations (NASA EOS Aura platform)
IONS-6 ozonesonde data for validation
The *ideal* strongly constrained 4D-Var problem uses a perfect model and perfect data

The ideal cost function:

\[
J(x_0) = \frac{1}{2} \left( x_0 - x_0^b \right)^T B_0^{-1} \left( x_0 - x_0^b \right) + \frac{1}{2} \sum_{k=0}^{N} \left( \mathcal{H}_k(x_k) - y_k \right)^T R_k^{-1} \left( \mathcal{H}_k(x_k) - y_k \right)
\]

The ideal strongly constrained 4D-Var problem is

\[
x_0^a = \arg \min_{x_0 \in \mathbb{R}^n} J(x_0)
\]

subject to:

\[
x_{k+1} = M_{k,k+1}(x_k), \quad k = 0, \ldots, N - 1.
\]
Practical 4D-Var problem uses an imperfect model and noisy data

Model is imperfect:

\[ \hat{x}_{k+1} = M_{k,k+1}(\hat{x}_k) + \Delta x_{k+1}, \quad k = 0, \ldots, N - 1. \]

Cost function includes imperfect data:

\[
\hat{J}(x_0) = \frac{1}{2} (x_0 - x_0^b)^T B_0^{-1} (x_0 - x_0^b) + \frac{1}{2} \sum_{k=0}^{N} (H_k(x_k) - y_k - \Delta y_k)^T R_k^{-1} (H_k(x_k) - y_k - \Delta y_k)
\]

The perturbed strongly constrained 4D-Var problem is

\[
\hat{x}_0^a = \arg \min_{x_0 \in \mathbb{R}^n} \hat{J}(x_0) \quad \text{subject to:} \quad \hat{x}_{k+1} = M_{k,k+1}(\hat{x}_k) + \Delta x_{k+1}, \quad k = 0, \ldots, N - 1.
\]
The solution obtained in an application obeys the optimality equations for the perturbed 4D-Var

The perturbed strongly constrained 4D-Var problem is

\[ \hat{x}_0^a = \arg \min_{x_0 \in \mathbb{R}^n} \hat{J}(x_0) \]

subject to:

\[ \hat{x}_{k+1} = \mathcal{M}_{k,k+1}(\hat{x}_k) + \Delta x_{k+1}, \quad k = 0, \ldots, N - 1. \]

The perturbed KKT conditions are:

**fwd:** \[ \Delta x_{k+1} = \hat{x}_{k+1} - \mathcal{M}_{k,k+1}(\hat{x}_k), \quad k = 0, \ldots, N - 1 \]

**adj:** \[ -H_N^T R_N^{-1} \Delta y_N = \lambda_N - H_N^T R_N^{-1} (\mathcal{H}_N(\hat{x}_N^a) - y_N), \]

\[ -H_k^T R_k^{-1} \Delta y_k = \lambda_k - M_{k,k+1}^T \lambda_{k+1} - H_k^T R_k^{-1} (\mathcal{H}_k(\hat{x}_k^a) - y_k), \quad k = N - 1, \ldots, 0 \]

**opt:** \[ 0 = B_0^{-1}(\hat{x}_0^a - x_0^b) + \lambda_0. \]
Deterministic view: estimate the error in an aspect of interest of the inference solution

Aspect of interest (a.o.i.): $\mathcal{E} : \mathbb{R}^n \rightarrow \mathbb{R}$

$x_0 \rightarrow \mathcal{E}(x_0)$

Want: error in a.o.i.: $\Delta \mathcal{E} = \mathcal{E}(\hat{x}_0^a) - \mathcal{E}(x_0^a)$

Approach: express error in terms of weighted optimality residuals:

$\Delta \mathcal{E} \approx \langle \nu, \Delta F \rangle + \langle \mu, \Delta A \rangle + \langle \zeta, \Delta O \rangle$

$\quad = \langle \nu, \Delta x \rangle + \langle \mu, \Delta y \rangle$

Solution ingredients: TLM, ADJ, SOA, large linear systems
Shallow water equations on the sphere provide a large-scale model of the atmosphere

\[ \begin{align*}
\frac{\partial u}{\partial t} + \frac{1}{a} \cos \phi \frac{\partial u}{\partial x} + v \cos \phi \frac{\partial u}{\partial \phi} + f u \tan \phi \frac{\partial h}{\partial \phi} + g a \cos \phi \frac{\partial h}{\partial \phi} &= 0, \\
\frac{\partial v}{\partial t} + \frac{1}{a} \cos \phi \frac{\partial v}{\partial x} + v \cos \phi \frac{\partial v}{\partial \phi} + f v \tan \phi \frac{\partial h}{\partial \phi} + g a \cos \phi \frac{\partial h}{\partial \phi} &= 0, \\
\frac{\partial h}{\partial \phi} + \frac{1}{a} \cos \phi \frac{\partial (hu)}{\partial \phi} + \frac{\partial d}{\partial \phi} (hv \cos \phi) &= 0,
\end{align*} \]

Here, \( f \) is the Coriolis parameter given by \( f = 2 \Omega \sin \phi \), where \( \Omega \) is the angular speed of the rotation of the earth, \( h \) is the height of the homogeneous atmosphere, \( u \) and \( v \) are the zonal and meridional wind components, respectively, \( \phi \) and \( \lambda \) are the latitudinal and longitudinal directions, respectively, \( a \) is the radius of the earth and \( g \) is the gravitational constant. The space discretization is performed using the unstaggered Turkel-Zwas scheme [1]. The number of longitudinal nodes to perform the discretization is \( n_{\text{lon}} \) and the number of latitudinal nodes for the discretization is \( n_{\text{lat}} \). We obtain the following ODE:

\[ y(t) = f(t, y(t)) \quad y(t_0) = y_0, \quad t = [0, 3600]. \]

In (55) the zonal wind, meridional wind and the height is combined into the vector \( y \). Hence the size of \( y \) is \( 3 \times n_{\text{lat}} \times n_{\text{lon}} \). With the ODE as the model, we have some observations. The cost function is constructed based on the mismatch between the observations and the values obtained by solving the model. The cost function can be written as:

\[ J(y_0) = \frac{1}{2} \sum_i H(y_i) y_{\text{obs}}^T R_i H(y_i) y_{\text{obs}} \]

For our problem we use \( R_i \) as a diagonal matrix to scale such that all the \( y_i \)'s are approximately in the same range. Now the inverse problem can be formulated as:

\[ \min_{y_0} J(y_0) \quad \text{subject to } (55). \]

Our quantity of interest is:

\[ E = \frac{1}{N_{\text{grid}}} \sum_{i=1}^{N_{\text{grid}}} y \left( t_0 \right). \]

or

\[ E = \frac{1}{\text{area sphere}} \int_{\text{sphere}} h^a(t_0, x) \, dx \]

or

\[ E = \frac{1}{N_{\text{grid}}} \sum_{i=1}^{N_{\text{grid}}} y \left( t_0 \right) y_{\text{opt}}^T \]

Quantity of interest:
The methodology can parse the overall impact of the errors in individual data points (one realization)

- Data errors: normal,
- std ~2% (of max)

- Data error impact:
  Actual: 54.70
  Estimated: 57.26
Similarly, the contributions of model errors in $h$, $u$, and $v$ at each grid point are assessed (discrete time):

- Model errors (discrete):
  - Fixed (bias), std $\sim$2%(max)
- Model error impact:
  - Actual: 1.97
  - Estimated: 2.96
Probabilistic view: the data/model errors are described by probability densities.

- Errors vs. fine grid, small time step
- 216 error snapshots at different times
- Fit distributions: Bayes info criterion
- Errors ~ Gaussian process
- Inter-grid correlations ~ Bessel 1st kind
Probabilistic view: aposteriori estimates of the moments of the error in the a.o.i.

Estimates of the mean and variance of the impact of model errors:

\[
\Delta \mathcal{E}_\text{mod} = - \sum_{k=1}^{N} \nu_k^T \Delta \mathbf{x}_k \quad \text{(deterministic)},
\]

\[
\mathbb{E}[\Delta \mathcal{E}_\text{mod}^{\text{stat}}] \approx - \sum_{k=1}^{N} \bar{\nu}_k^T \bar{\Delta \mathbf{x}}_k,
\]

\[
\text{Var}\left(\Delta \mathcal{E}_\text{mod}^{\text{stat}}\right) \approx \sum_{k=1}^{N} \sum_{\ell=1}^{N} \bar{\nu}_k^T \text{Cov}\left(\Delta \mathbf{x}_k, \Delta \mathbf{x}_\ell\right) \bar{\nu}_\ell.
\]

<table>
<thead>
<tr>
<th></th>
<th>( \mathbb{E}[\Delta \mathcal{E}_{\text{obs}}] )</th>
<th>( \text{Var}\left[\Delta \mathcal{E}_{\text{obs}}\right] )</th>
<th>( \mathbb{E}[\Delta \mathcal{E}_{\text{mod}}] )</th>
<th>( \text{Var}\left[\Delta \mathcal{E}_{\text{mod}}\right] )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimates</td>
<td>0.00</td>
<td>2.87</td>
<td>1.21</td>
<td>0.053</td>
</tr>
<tr>
<td>Ensemble (errors)</td>
<td>0.105</td>
<td>2.53</td>
<td>1.17</td>
<td>0.080</td>
</tr>
</tbody>
</table>
New questions: quantify data impact, and define optimal sensor network configurations

$$u_{opt} = \arg\min_u \Psi(x^a_v) \quad \text{subject to:} \quad \begin{cases} x^a_0 = \arg\min_{x_0} J(x_0), \\ x^a_v = M_{t_0 \rightarrow t_v}(x^a_0). \end{cases}$$

[Cioaca & Sandu, 2013]
Faulty sensors are not visible from the inference solution, but can be detected via the error impact.

4D-Var h-increment is smooth despite two faulty sensors

The two faulty sensors have a large, negative impact.

[Cioaca & Sandu, 2012]
Conclusions

- Completed framework to characterize impact of data and model errors on DDDAS ML inference results
- Both deterministic and probabilistic approaches are considered
- Formulated and solved the data weights and sensor (re)location problems as two-level optimization problems

Current and future work:
- Characterize data and model errors impact for ensemble based inference approaches
- Quantify error impact on sensor configuration
DyDESS 2014

Dynamic Data-driven Environmental Systems Science Conference

Nov 5 – 7, Massachusetts Institute of Technology, Cambridge, MA

Call for Papers

Submission Deadline: August 15, 2014 (dydess_submit@mit.edu)
Invitations/Acceptance: September 15, 2014
Registration & Camera Ready: October 15, 2014
Conference Dates: Nov 5 -7, 2014
Location: MIT, Cambridge, MA.
URL: dydess.mit.edu

http://www.dydess.mit.edu