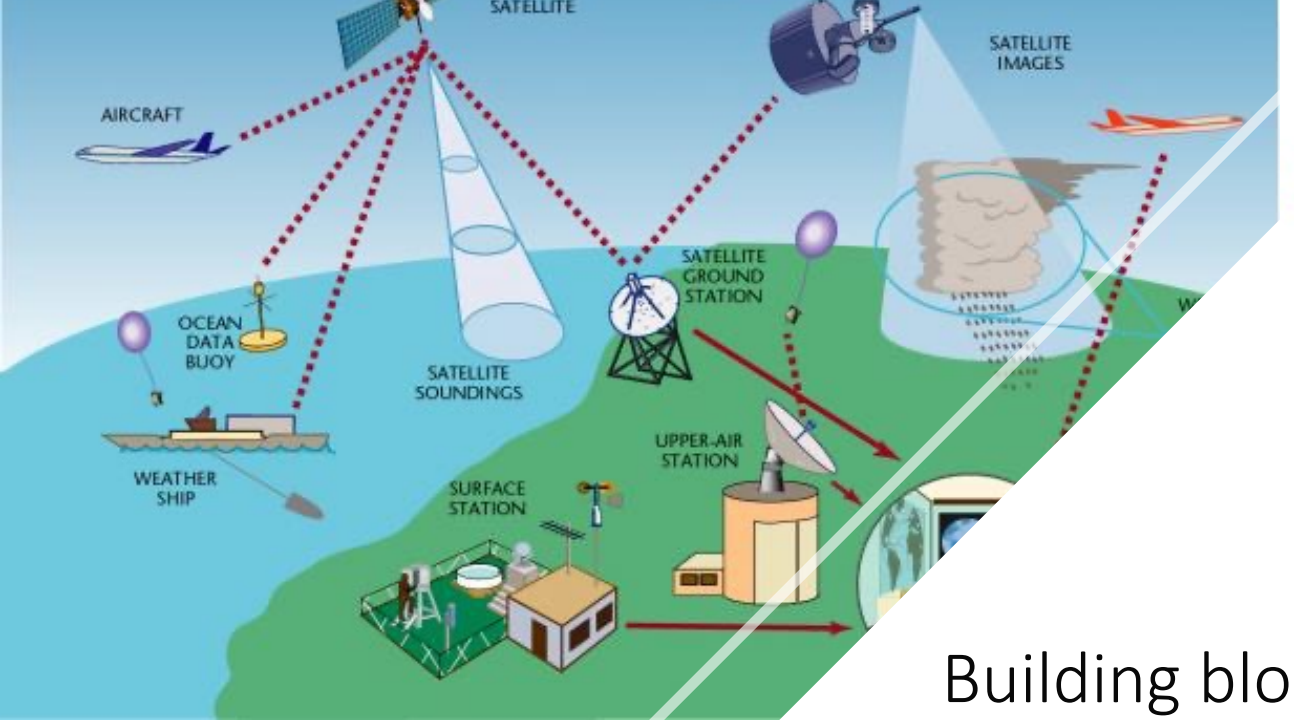
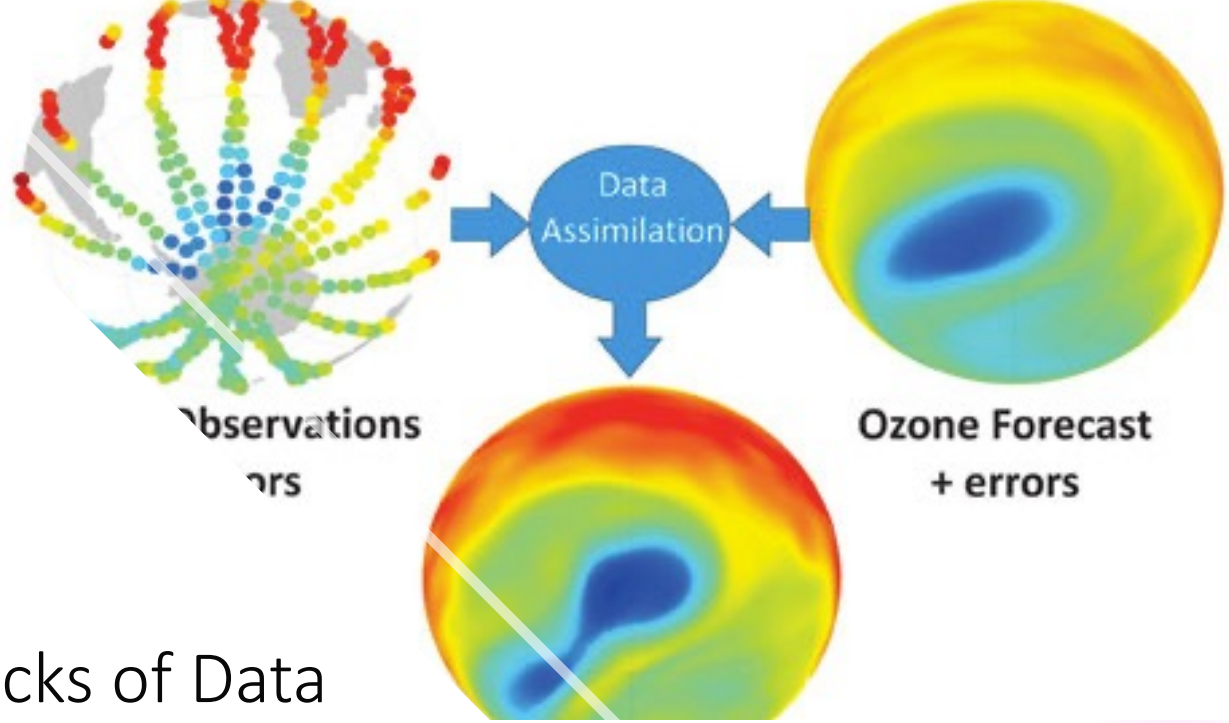
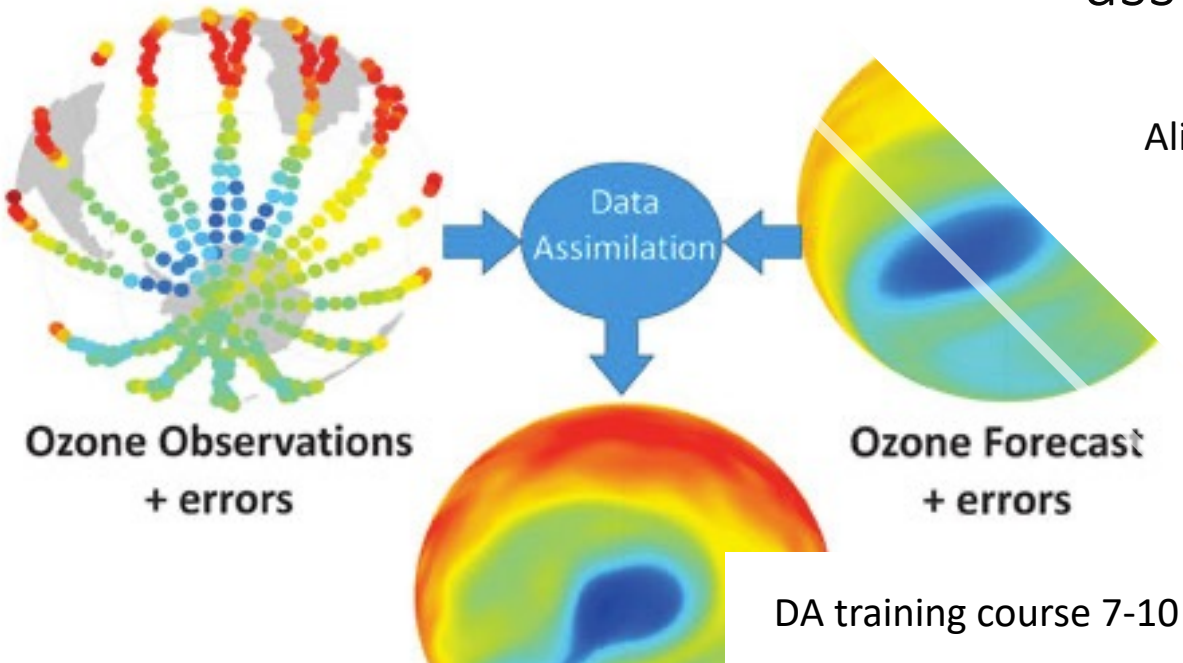
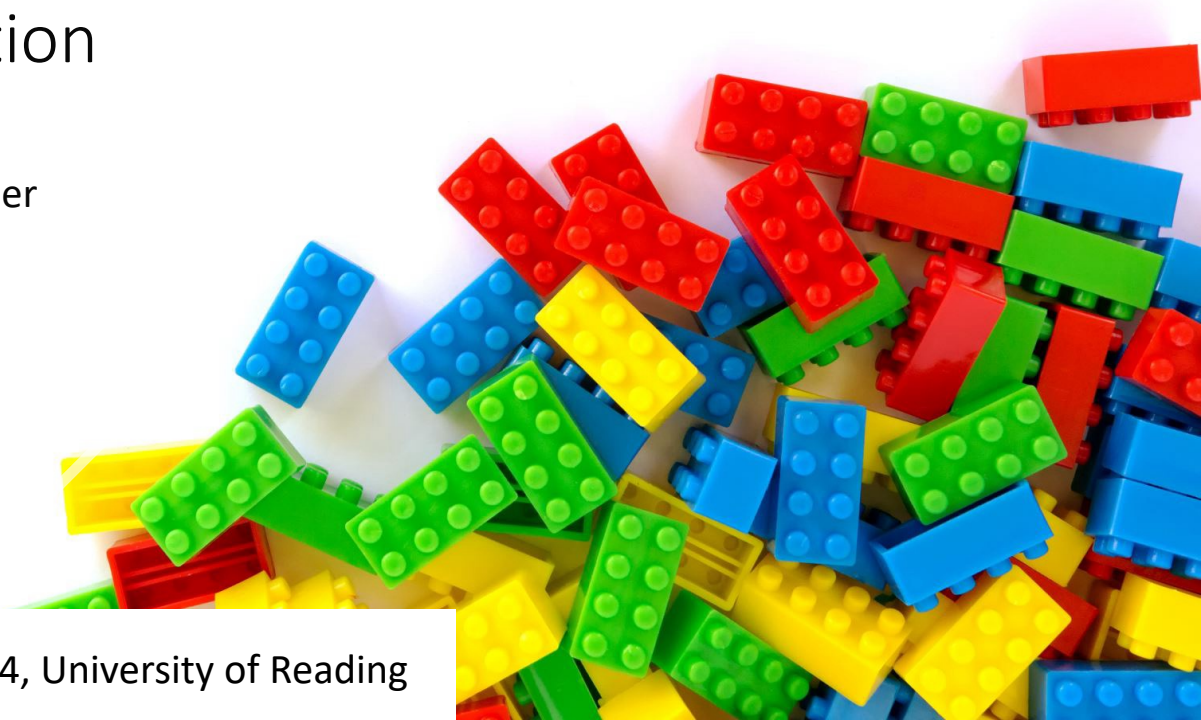


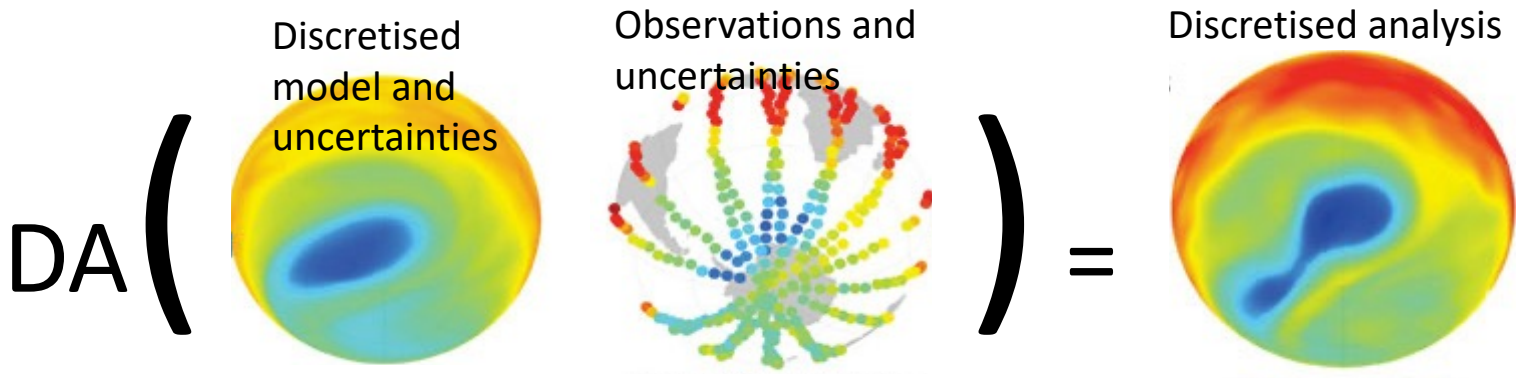


Building blocks of Data assimilation

Alison Fowler

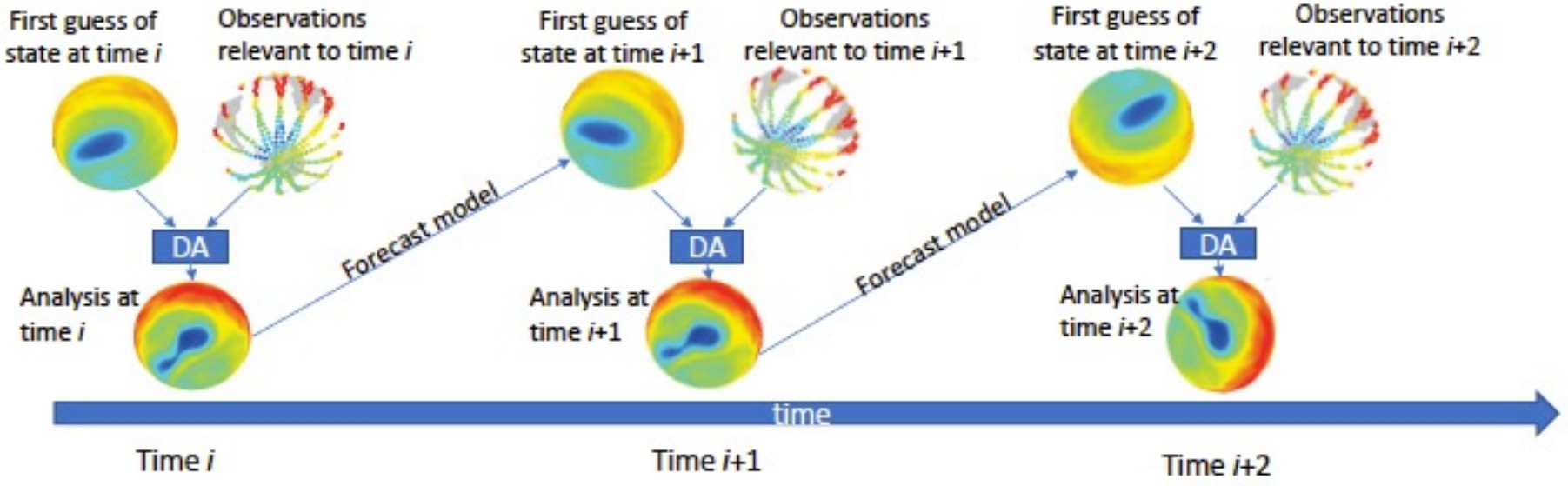


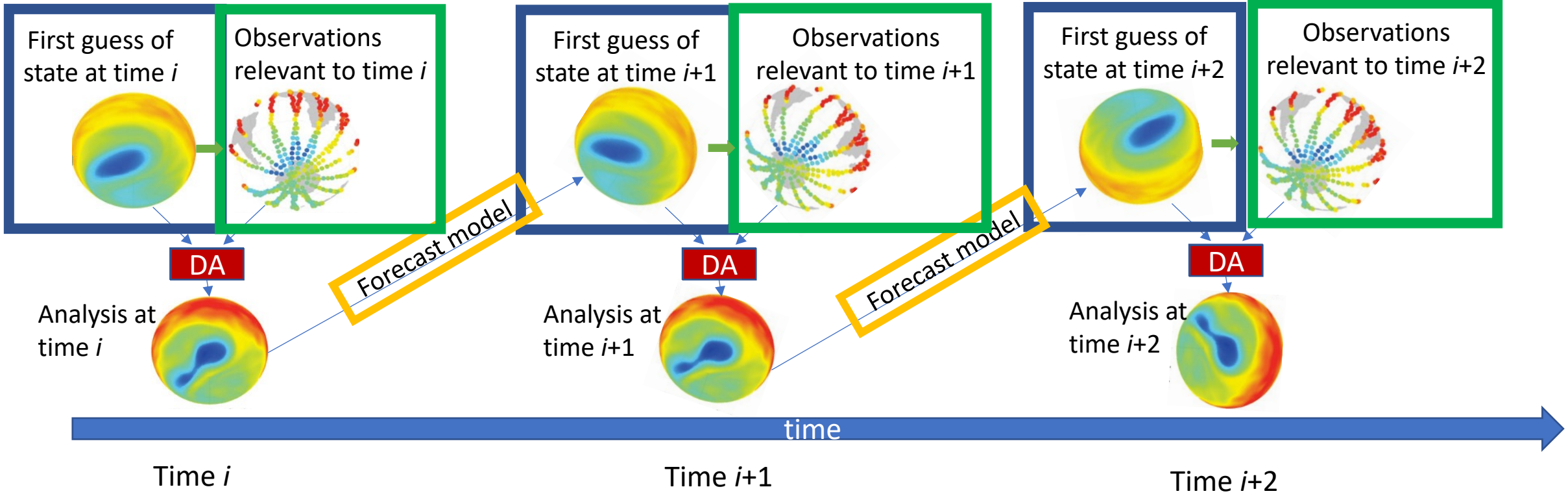




The DA problem is to combine prior knowledge and relevant observations to give an updated estimate of the current atmosphere/oceans/land surface etc. This is then used to initialize a forecast.

The assimilation is cycled in time so that information from the observations can be continuously fed into the forecast.





The aim of this lecture is to explain the DA ingredients: what they represent and where they come from.

DA ingredients:

- First guess of the model state known as the background
- The background uncertainty
- Observations
- The observation uncertainty
- A mapping form the model grid and variables to the observation variables and locations, known as the observation operator
- A forecast model
- Data assimilation methodology

Data assimilation methodology

To understand why each of the DA ingredients are important, let's first look at how they are used. This will depend upon the DA method implemented.

Different DA methods will be described over the following days.

Here we will give the basic idea of the **Kalman equations** that form the basis of many DA methods. This assumes that the **errors** in the observation and background are **Gaussian** and **unbiased**.

The analysis equation:

$$\mathbf{x}_{\mathrm{a}}^{i} = \mathbf{x}_{\mathrm{b}}^{i} + \mathbf{K}^{i}\left(\mathbf{y}^{i} - h\left(\mathbf{x}_{\mathrm{b}}^{i}\right)\right) = \text{first guess} + \text{update},$$

i.e. the analysis is a weighted combination of the background ($\boldsymbol{x}_{b}^{i}$) and the observations ($\mathbf{y}^{i}$). i is a time index indicating the assimilation cycle.

The Kalman gain matrix:

$$\mathbf{K}^{i} = \mathbf{B}^{i}\ (\mathbf{H}^{i})^{T}(\mathbf{H}^{i}\mathbf{B}^{i}\ (\mathbf{H}^{i})^{T} + \mathbf{R}^{i})^{-1},$$

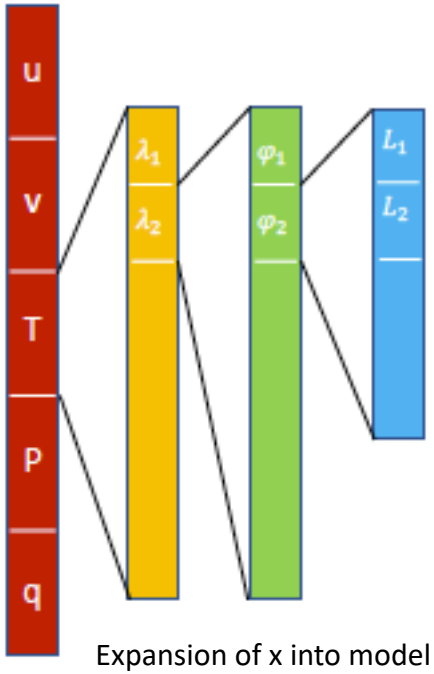
a function of the background and observation error covariance matrices ($\mathbf{B}$ and $\mathbf{R}$) and the linearised observation operator ($\mathbf{H}$).

These can be derived from finding the state $\mathbf{x}$ with the minimum error variance or equivalently (in the case of Gaussian errors) the maximum a-posteriori probability.

Vector notation

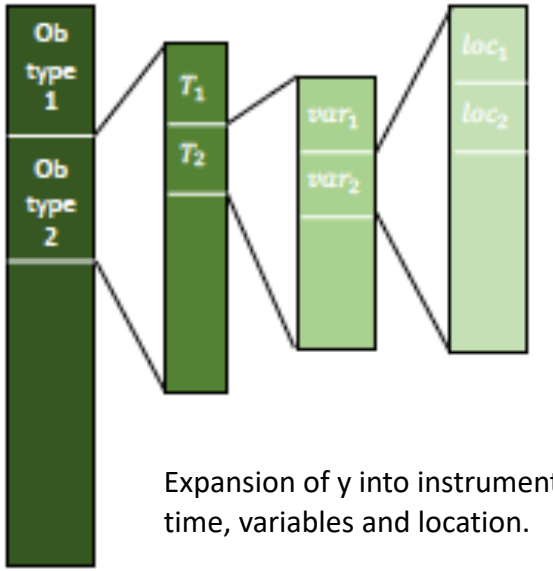
$\mathbf{x}^i \in \mathbb{R}^n$ is a column vector containing all the model variables that you wish to update via data assimilation at time i. In NWP n is typically of the order 10^8.

$\mathbf{y}^i \in \mathbb{R}^p$ is a column vector containing all the relevant observations in updating $\mathbf{x}^i$. This will include observations from different instruments, measuring different variables, and possibly be at different times. In NWP p is typically of the order 10^6.



Expansion of x into model variables, latitude, longitude and model levels.

The observation operator, $h^i: \mathbb{R}^n \rightarrow \mathbb{R}^p$ is a vector mapping from state to observation space



Expansion of y into instruments, time, variables and location.

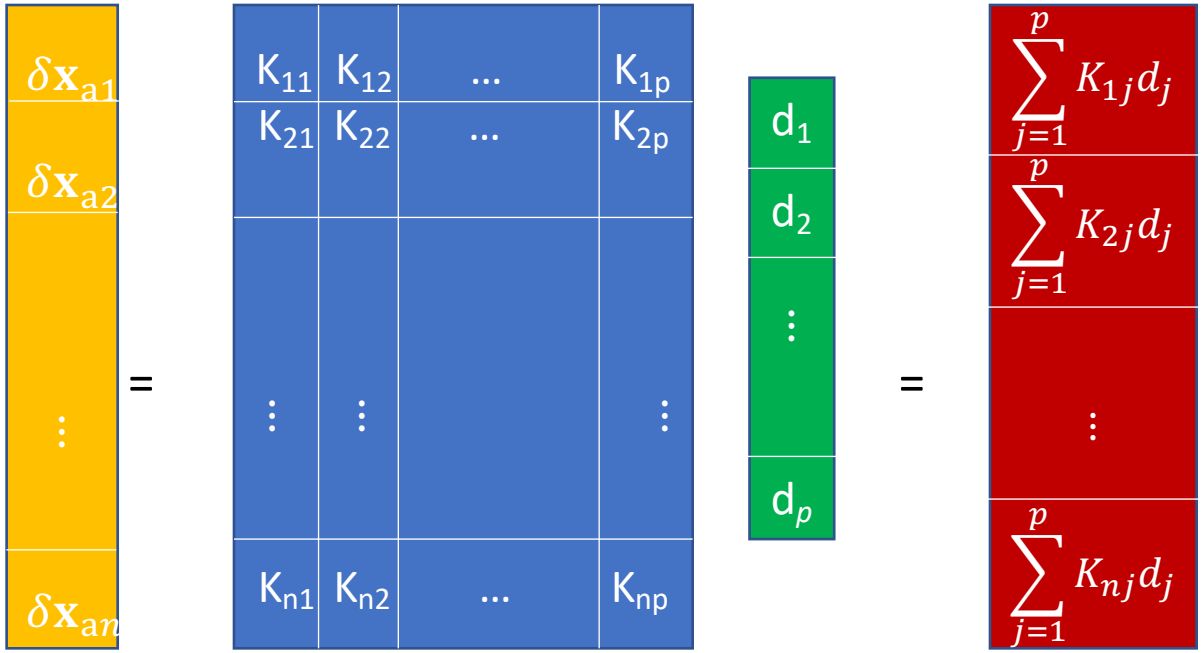
Matrix notation

The update to the background is given as $\mathbf{x}_\mathrm{a} = \mathbf{x}_\mathrm{b} + \mathbf{K}(\mathbf{y} - h(\mathbf{x}_\mathrm{b}))$

or $\delta\mathbf{x}_\mathrm{a} = \mathbf{K}\,\mathbf{d},$

where $\delta\mathbf{x}_\mathrm{a} \in \mathbb{R}^n$ is the analysis increment and $\mathbf{d} \in \mathbb{R}^p$ is the innovation. (I've dropped the time index as was getting a bit unwieldly!).

The Kalman gain then has dimensions $\mathbf{K} \in \mathbb{R}^{?\times?}$



Each element of $\delta\mathbf{x}_a$ is a linear combination of the innovations, with weightings given by the rows of $\mathbf{K}$.

$$\mathbf{K} = \mathbf{B}\mathbf{H}^\mathrm{T}(\mathbf{H}\mathbf{B}\mathbf{H}^\mathrm{T} + \mathbf{R})^{-1}$$

The background and its uncertainty

The background, to be updated by DA, often comes from the forecast initialized from an analysis from the previous assimilation time.

$$\mathbf{x}_{\mathrm{b}}^{i} = M(\mathbf{x}_{\mathrm{a}}^{i-1}, parameters, forcing)$$

The background error is defined as $\boldsymbol{\varepsilon}_{\mathrm{b}} = \mathbf{x}_{\mathrm{b}} - \mathbf{x}_{truth}$

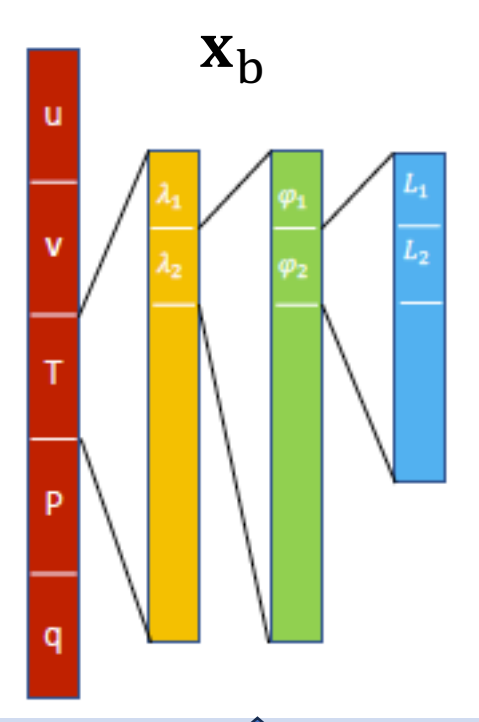
Sources of error in the background include

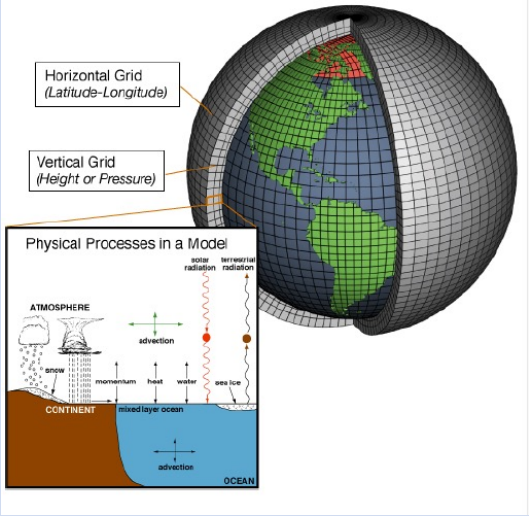
- Initial conditions (i.e. the previous analysis)
- Intrinsic model error growth
- The model equations
- Discretisations of model equations
- Parmeterisations
- External forcing

These different sources of error can make it very difficult to quantify the background error statistics. However, we can make approximations to allow us to give a realistic description.

The first approximation is **to assume that the errors are unbiased and Gaussian** (or alternatively we have been able to remove biases and we have transformed to variables that have Gaussian errors).

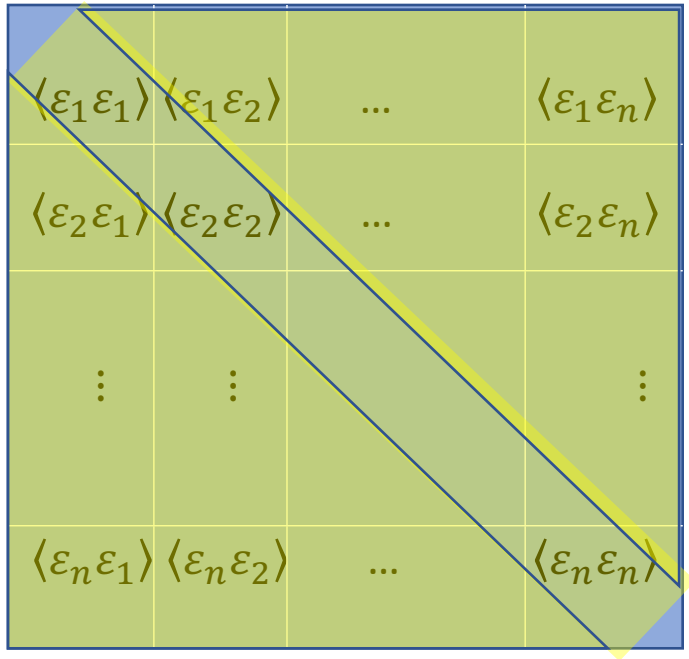
This leave us to define the background error covariance matrix – the B matrix!





Error covariance matrices

If the errors in the background and observations are unbiased and Gaussian, then their statistics can be described solely by their respective error covariance matrices. $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{C})$



An error covariance matrix is given by the expectation of the outer product of the errors.

$$\mathbf{C} = \left\langle \boldsymbol{\varepsilon} \boldsymbol{\varepsilon}^{\mathrm{T}} \right\rangle, \text{ where e.g. } \boldsymbol{\varepsilon} = \boldsymbol{x}_b - \boldsymbol{x}_{truth}$$

Note we have not removed the mean because we assume the errors are unbiased i.e. have a zero mean.

The error variances are given by the diagonal elements, $\sigma_i^2 = \langle \varepsilon_i \varepsilon_i \rangle$.

The error covariances are given by the off-diagonal elements.

By definition an error covariance matrix is symmetric, $\left\langle \varepsilon_i \varepsilon_j \right\rangle = \left\langle \varepsilon_j \varepsilon_i \right\rangle$ and positive definite.

The error correlations are the covariances normalized by the respective error standard deviations:

$$corr_{ij} = (\mathbf{C}_{ij})/(\sigma_i \sigma_j)$$

$$corr_{ij} = \frac{\left\langle \varepsilon_i \varepsilon_j \right\rangle}{\sqrt{\langle \varepsilon_i \varepsilon_i \rangle} \sqrt{\left\langle \varepsilon_j \varepsilon_j \right\rangle}}$$

$$\mathbf{x}_\mathrm{a} = \mathbf{x}_\mathrm{b} + \mathbf{K}(\mathbf{y} - h(\mathbf{x}_\mathrm{b}))$$

The B matrix

$\mathbf{B} \in \mathbb{R}^{n \times n}$ has n^2 elements describing how the errors in each model variable covary spatially as well as how errors in different variables covary.

The importance of its structure is seen in the expression for the Kalman gain

$$\mathbf{K}^i = \mathbf{B}^i \, (\mathbf{H}^i)^T (\mathbf{H}^i \mathbf{B}^i \, (\mathbf{H}^i)^T + \mathbf{R}^i)^{-1}.$$

Recall this governs how the innovations are linearly weighted to update the background and provide the analysis. As $\mathbf{B}$ is the final operator, the analysis increments will lie in the subspace spanned by $\mathbf{B}$.

The structure of $\mathbf{B}$ is essential for defining how information in observations is spread to other locations and variables. It is also important for minimizing the risk of the analysis increments resulting in an analysis that is inconsistent with the model equations.

The B matrix- example of a single observation experiment

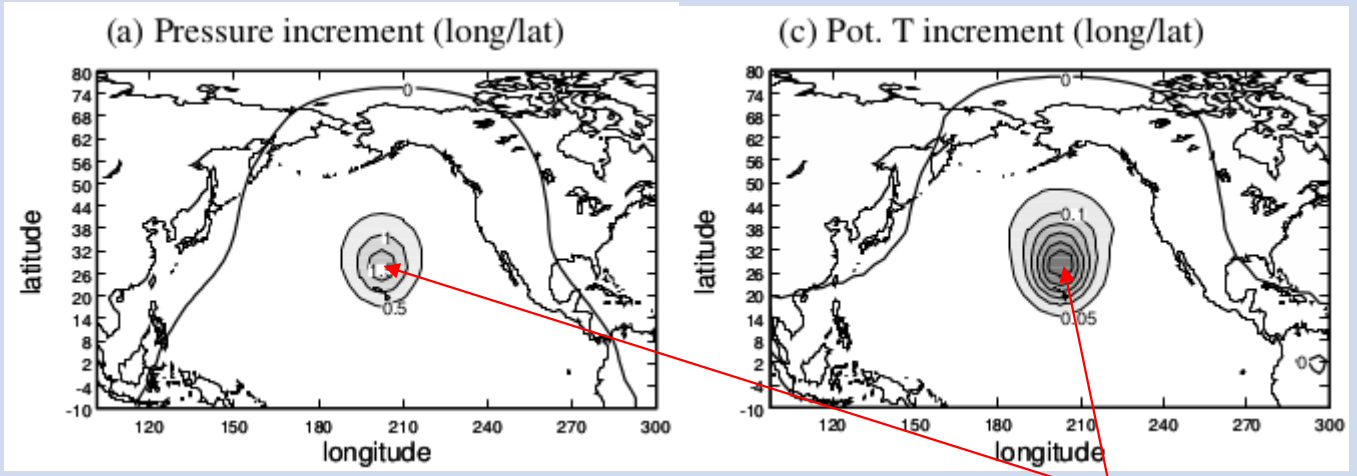
Let a single observation observe the i^{th} state variable such that

$$\mathbf{H} = (0 \quad \ldots \quad 1 \quad \cdots \quad 0),$$

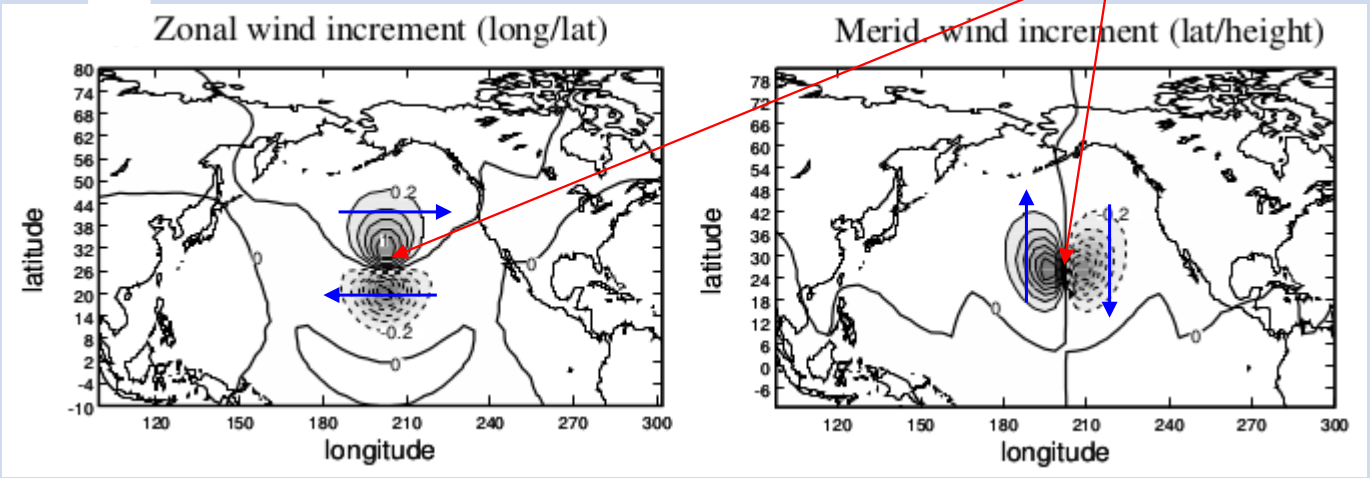
$$\mathbf{R} = \sigma_o^2 \text{ and } \mathbf{HBH}^{\mathrm{T}} = B_{ii}.$$

Then $\mathbf{x}_{\mathrm{a}} = \boldsymbol{x}_{\mathrm{b}} + \begin{pmatrix} B_{1i} \\ \vdots \\ B_{ii} \\ \vdots \\ B_{ni} \end{pmatrix} \frac{y - x_{bi}}{\sigma_o^2 + B_{ii}}.$

$$\underbrace{\mathbf{BH}^{\mathrm{T}}\left(\mathbf{HBH}^{\mathrm{T}} + \mathbf{R}\right)^{-1}}_{\mathbf{K}} d$$



Structure function *i* (in this case *i* is the pressure field at this position)



In this case the wind part of the structure function is in geostrophic balance with the pressure

Figure supplied by Ross Bannister

The B matrix

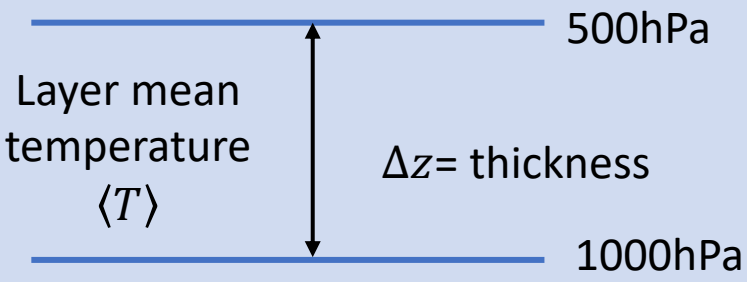
Due to its size, $\mathbf{B}$ can rarely be represented explicitly.

$\mathbf{B}$ can be simplified by enforcing balance relationships, the modelling of the spatial correlations, and making assumptions about how it changes in time. In this way the number of parameters needed to describe $\mathbf{B}$ is reduced.

For example (See Bannister 2008b for a comprehensive review)

- ‘Inverse Laplacians’
- Diffusion operators (commonly used in the ocean)
- Recursive filters
- Spectral and wavelet methods
- Exploitation of physics (e.g. geostrophic balance)
- Control variable transforms (transform to a space where B is simpler e.g. diagonal)

Control variable transform



Define a transform matrix $\mathbf{U}$ such that $\mathbf{U}\mathbf{U}^{\mathrm{T}} \approx \mathbf{B}$. The idea is then to perform the assimilation for the variable $\delta\mathbf{v}$ instead of $\delta\mathbf{x}$, where $\delta\mathbf{x} = \mathbf{U}\delta\mathbf{v}$. The error covariances for $\delta\mathbf{v}$ will, by definition, be the identity, greatly simplifying the DA algorithms. After the analysis for $\delta\mathbf{v}$ is found it can be transformed back to the space of the modelled variables and the analysis updates will implicitly account for the covariances of the original **B**.

Example, modelling hydrostatic balance:

- Let $\delta\mathbf{x} = \begin{pmatrix} \delta\langle T \rangle \\ \delta\Delta z \end{pmatrix}$ be our model variables that we wish to update.
- Split $\delta\Delta z$ into a part balanced with the mean temperature (L$\delta\langle T \rangle$) and an unbalanced part ($\delta\Delta z_{unbal}$), where $L = \frac{R}{10g} ln \frac{z_i}{z_{i-1}}$.
- Define our control variables as $\delta\boldsymbol{v}$= $\begin{pmatrix} \delta v_{bal} \\ \delta v_{unbal} \end{pmatrix}$ with $\begin{pmatrix} \delta\langle T \rangle \\ \delta\Delta z \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ L & 1 \end{pmatrix} \begin{pmatrix} \sigma_{bal} & 0 \\ 0 & \sigma_{unbal} \end{pmatrix} \begin{pmatrix} \delta v_{bal} \\ \delta v_{unbal} \end{pmatrix}$
- The implied covariances are then $\mathbf{B} = \mathbf{U}\mathbf{U}^{\mathrm{T}} = \begin{pmatrix} \sigma_{bal}^2 & \sigma_{bal}^2 L \\ \sigma_{bal}^2 L & \sigma_{bal}^2 L^2 + \sigma_{unbal}^2 \end{pmatrix}$.
- Observations of $\langle T \rangle$ or Δz will now update the analysis of the other variable in a physically consistent way.

U describes the hydrostatic balances scaled by the error standard deviations of the control variables. The key is there are no correlations in the errors of these variables.

Estimating the B-matrix

Even with simplifications, it is necessary to estimate the key features of the B matrix. This may be done in a few ways each acknowledging the sources of the errors in the background in different ways (See Bannister 2008a for a comprehensive review).

For example

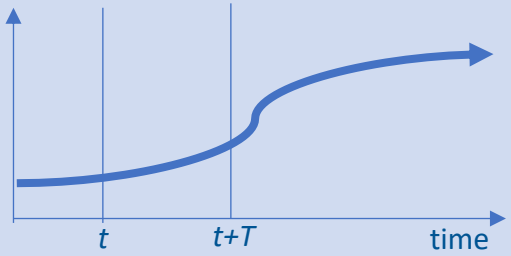
- Forecast differences
- Analysis of innovations, e.g. H-L method or consistency diagnostics
- Ensembles

Each of these methods estimates the error covariance matrix using a sample of the approximated background errors.

$B_{ij} = \langle \varepsilon_i \varepsilon_j \rangle \approx \frac{1}{N_{ij}-1} \sum_{k=1}^{N_{ij}} \varepsilon_{i,k} \varepsilon_{j,k}$, where N_{ij} is the sample size available for estimating B_{ij}.

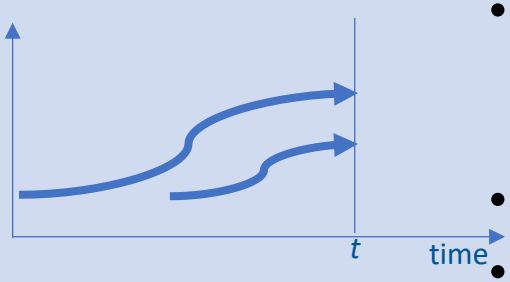
The difference between the methods is how the samples of the errors, ε , are obtained.

Estimating the B-matrix: Forecast differences



The background error can be approximated by comparing two forecasts:

- Canadian quick method Polavarapu et al. (2005):
  - Compares forecast valid at different times, $\boldsymbol{\varepsilon} \approx (\mathbf{x}(t+T) - \mathbf{x}(t))/\sqrt{2}$, where T is fixed.
  - The sample population is given by subsampling from one long forecast run.
  - Has advantage of being able to provide estimates of $\mathbf{B}$ prior to assimilation system being available.
- NMC method (Parrish and Derber (1992)):
  - Compares pairs of lagged forecasts valid for the same time but initialized from different background (or analysis) fields, e.g. the difference between a 30 and 6 hour forecast valid at time t, $\boldsymbol{\varepsilon} \approx (\mathbf{x}_{30}(t) - \mathbf{x}_{6}(t))/\sqrt{2}$.
  - The sample population is given by cycling the DA system.
  - Can have problems in poorly observed regions where the forecasts do not differ much.



Both estimated matrices can expect to need some scaling to represent the error in the background (often a 6 hour forecast).

Estimating the B-matrix: Analysis of innovations

The innovation can be written in terms of the errors in the observations and background transformed to observation space:

$$\mathbf{d} = \mathbf{y} - h(\mathbf{x}_b) = \mathbf{y} - h(\mathbf{x}_{truth}) + h(\mathbf{x}_{truth}) - \ h(\mathbf{x}_\mathrm{b}\ ) = \boldsymbol{\eta} - \mathbf{H}\boldsymbol{\varepsilon}$$

The covariance of the innovations is then $\mathbf{D} = \langle \mathbf{d}\mathbf{d}^\mathrm{T} \rangle = \mathbf{R} + \mathbf{H}\mathbf{B}\mathbf{H}^\mathrm{T}$, assuming that the errors in the observations and background are uncorrelated, and the errors are unbiased.

An estimate of $\mathbf{D}$ is then possible given a sample population of innovations from a cycled DA system.

$\mathbf{H}\mathbf{B}\mathbf{H}^\mathrm{T}$ can then be isolated from $\mathbf{D}$ by

1. making assumptions about $\mathbf{R}$, e.g. observation errors are uncorrelated, and assuming the background covariances are isotropic and homogeneous e.g. Hollingworth and Lönnberg, 1986
2. Or by using additional information in the analysis increments e.g. Desroziers et al. 2005 method.

Only gives information about B in the space of the observations (need to observe key model variables!).

Estimating the B-matrix: Ensembles

A sample of background errors can be provided by perturbing every known source of background uncertainty given assumptions about the distributions of these uncertainties.

This produces an ensemble of possible backgrounds all valid for the same time from which the background error is estimated, $\boldsymbol{\varepsilon} \approx \mathbf{x}_{\mathrm{b},k}(t) - \widehat{\mathbf{x}(t)}$, where $\mathbf{x}_{\mathrm{b},k}(t)$ is the kth ensemble member and $\widehat{\mathbf{x}(t)}$ is the ensemble mean.

The sample population is given by the ensemble, with the possibility of increasing the sample by including different assimilation times.

The sample of background errors can be compactly written in terms of the perturbation matrix $\mathbf{X}' \in \mathbb{R}^{n \times N}$, where N is the size of the ensemble.

$$\mathbf{X}' = \begin{pmatrix} \mathbf{x}_{\mathrm{b},0}(t) - \widehat{\mathbf{x}(t)} & \cdots & \mathbf{x}_{\mathrm{b},N-1}(t) - \widehat{\mathbf{x}(t)} \end{pmatrix}$$

$$\mathbf{B} \approx \frac{1}{N-1} \mathbf{X}' (\mathbf{X}')^{\mathrm{T}}$$

Estimating the B-matrix

What sources of error in the background can each method account for?

Source of error	Canadian Quick	NMC	Analysis of innovations	ensemble
Initial conditions (analysis)				
Intrinsic model error growth				
The model equations				
Discretisations of model equations				
Parmeterisations				
External forcing				

Estimating the B-matrix

What sources of error in the background can each method account for?

Source of error	Canadian Quick	NMC	Analysis of innovations	ensemble
Initial conditions (analysis)				
Intrinsic model error growth	✓			
The model equations				
Discretisations of model equations				
Parmeterisations				
External forcing				

Estimating the B-matrix

What sources of error in the background can each method account for?

Source of error	Canadian Quick	NMC	Analysis of innovations	ensemble
Initial conditions (analysis)		✓		
Intrinsic model error growth	✓	✓		
The model equations				
Discretisations of model equations				
Parmeterisations				
External forcing				

Estimating the B-matrix

What sources of error in the background can each method account for?

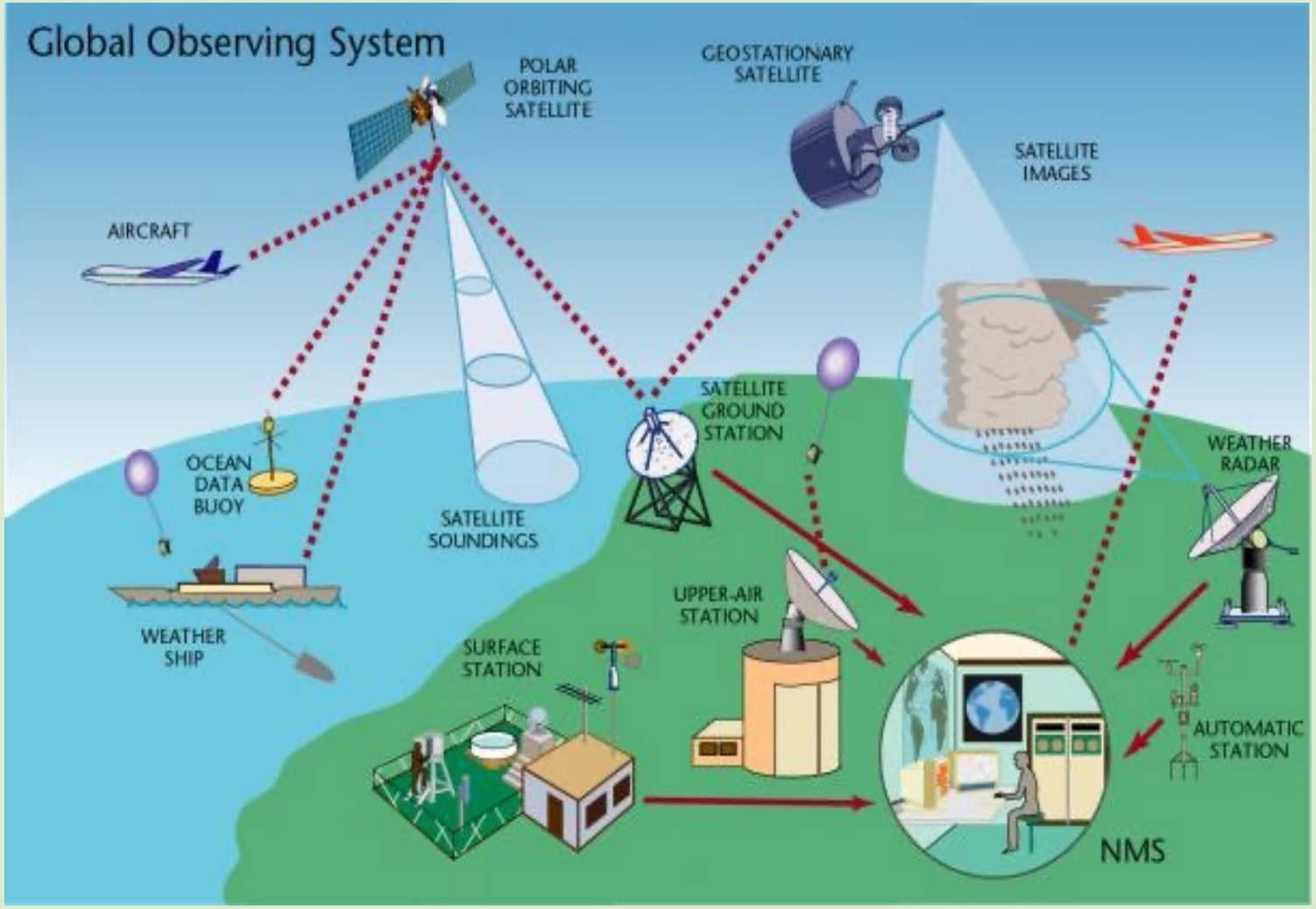
Source of error	Canadian Quick	NMC	Analysis of innovations	ensemble
Initial conditions (analysis)		✓	✓	
Intrinsic model error growth	✓	✓	✓	
The model equations			✓	
Discretisations of model equations			✓	
Parmeterisations			✓	
External forcing			✓	

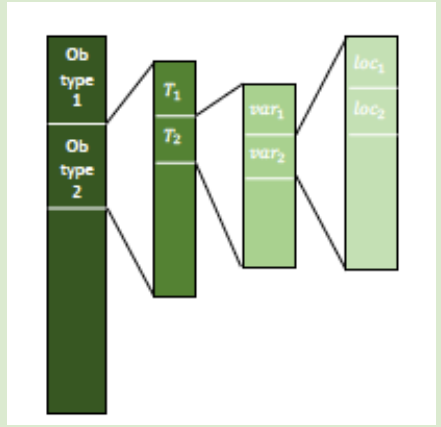
Estimating the B-matrix

What sources of error in the background can each method account for?

Source of error	Canadian Quick	NMC	Analysis of innovations	ensemble
Initial conditions (analysis)		✓	✓	✓
Intrinsic model error growth	✓	✓	✓	✓
The model equations			✓	depends
Discretisations of model equations			✓	depends
Parmeterisations			✓	depends
External forcing			✓	depends

The observations





Tens of millions of observations are assimilated globally in each cycle.

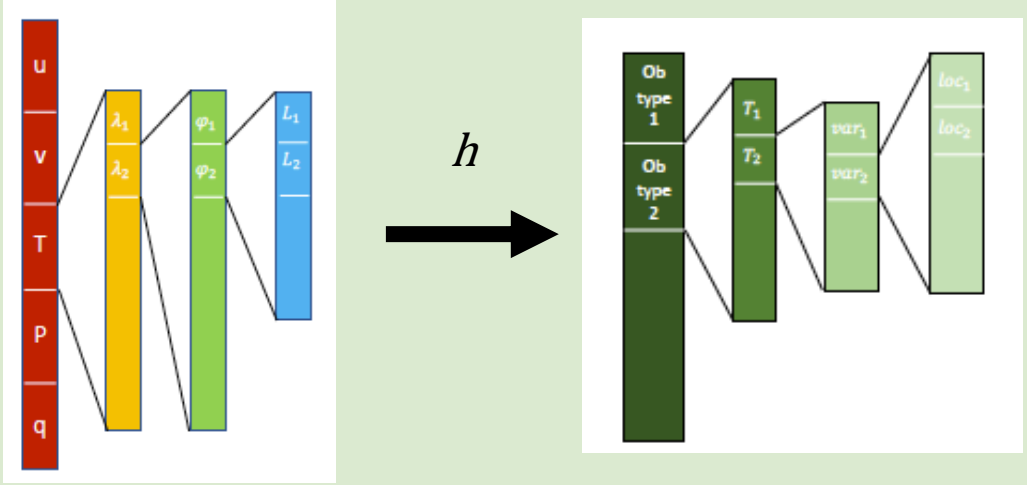
There are many different types of instruments.

- Over 90% of data comes from satellite radiances.

Before the observations are assimilated, they may be subject to

- Different levels of preprocessing,
- Quality control,
- adaptive thinning,
- Superobbing,
- bias correction (this could also be performed during assimilation)

The observation operator



To compare the modelled variables to the observations, the observation operator should account for:

- **Differences in location** of the observations to the the model grid (e.g. using interpolation)
- **Difference in variables** observed (e.g. using known physical or geometrical relationships, e.g. radiative transfer equations or projections from latitudal and londitudal winds to wind speed and direction).
- **Differences in scales** observed (e.g. if the observations observe larger scales than the modelled variables can use averaging, if the observations observer smaller scales than modelled this is trickier).
- **Differences in the timing** of the observations (e.g. using the dynamical model to evolve the initial model state to the time of the observations.)

If we are assuming that both the background and observation errors are Gaussian then this implies that the observation operator should be near-linear.

Observation errors

The errors in the observations are given by

$$\boldsymbol{\eta} = \mathbf{y} - h(\mathbf{x}_{\text{truth}})$$
$$= \underbrace{\mathbf{y} - h_{truth}(\mathbf{x}_{\text{truth}})}_{\boldsymbol{\eta}_{ob}} + \underbrace{h_{truth}(\mathbf{x}_{\text{truth}}) - h(\mathbf{x}_{\text{truth}})}_{\boldsymbol{\eta}_{h}}$$

The uncertainty in the observations is therefore given by a mixture of errors in the observation itself, $\boldsymbol{\eta}_{ob}$, and the errors in trying to compare the state to the observations, $\boldsymbol{\eta}_{h}$.

$\boldsymbol{\eta}_{ob}$ will include instrument errors and pre-processing errors (e.g. the effect of superobbing or converting to a more easily interpretable variable, can also include mistakes in cloud clearing)

$\boldsymbol{\eta}_{h}$ will include representation errors (e.g. observations of sub-grid scales) and errors in the observation operator.

Assuming the observation errors are unbiased and Gaussian then their statistics can be fully represented by the observation error covariance matrix – the R matrix!

The R matrix

It is often assumed that the R matrix is diagonal, greatly simplifying the DA implementation.

However, the complex sources of observation error mean that this assumption may be invalid.

If significant spatial error correlations are suspected, then a common approach is to thin the data so only uncorrelated data remains. This can result in a large proportion of observations being discarded that could contain useful information.

The importance of accounting for observation error correlations is increasingly being understood and are being included for a variety of observation types.

The R matrix: correlated errors

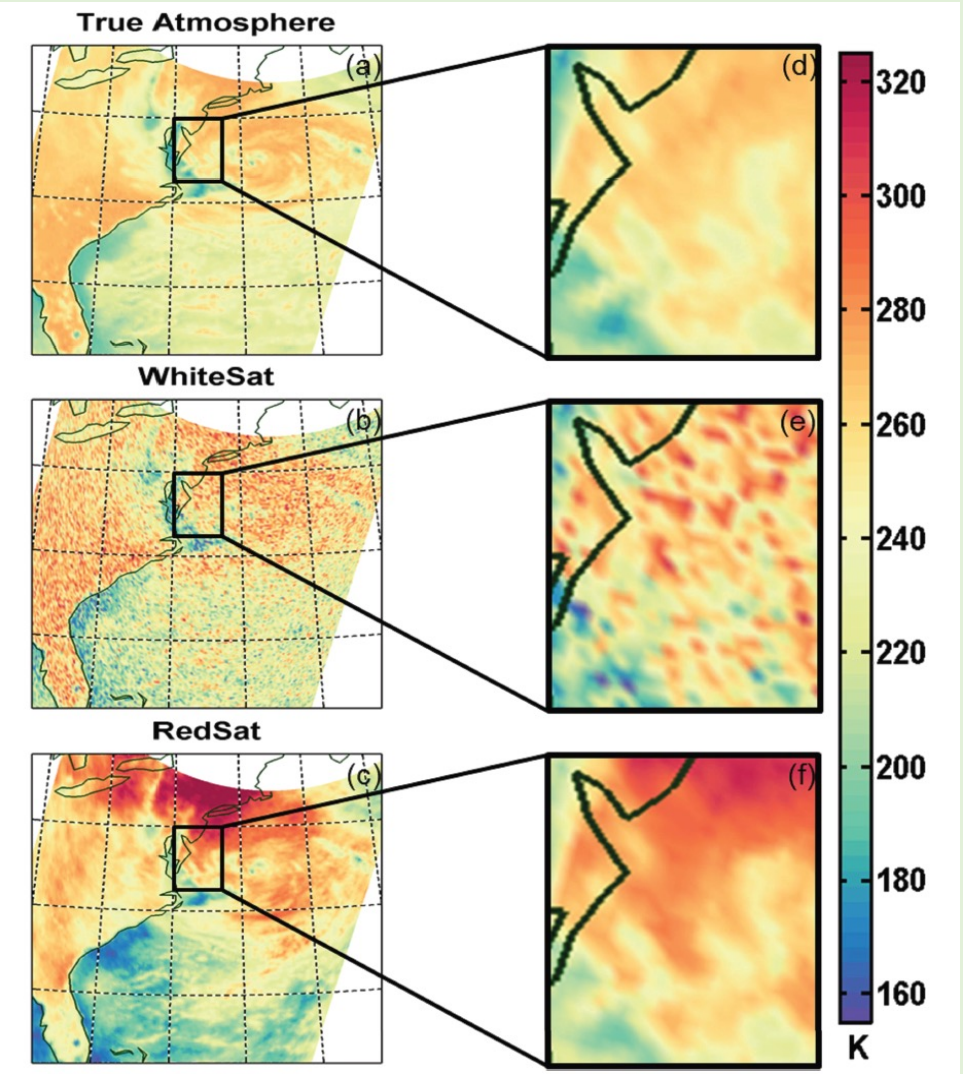


Figure 1. (a) A microwave satellite image of Hurricane *Sandy* on 24 October 2012, which is treated as truth. (b) Panel (a) plus white (uncorrelated) noise; (c) panel (a) plus red (spatially correlated) noise. (d)–(f) Detail views of (a)–(c), respectively. The colour scale for all panels is brightness temperature in Kelvin. This figure is for illustrative purposes.

Spatial error correlations imply that the observations contain less information about the mean (large scale) fields but more about gradients (small scales).

Rainwater et al. 2015, QJRMS

Estimating the R matrix

There are two main approaches to estimating $\mathbf{R}$:

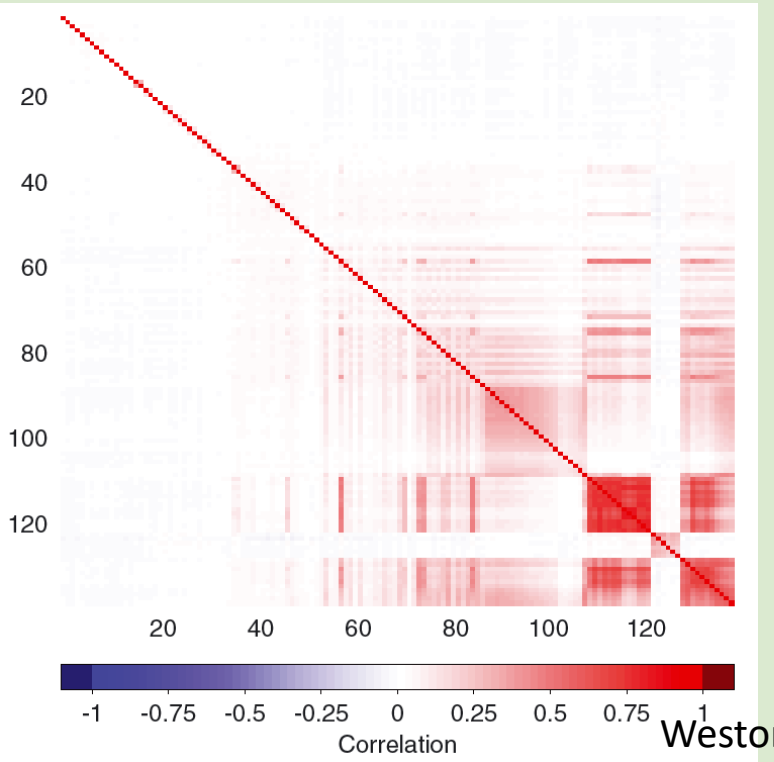
- The metrological approach: The accumulative effect of each source of error is meticulously traced from its origin to the final observation assimilated. This has analogies with the ensemble approach for estimating $\mathbf{B}$.
- The analysis of innovations: We have already seen that the covariance of the innovations gives the sum of $\mathbf{R}$ and $\mathbf{HBH}^{\mathrm{T}}$. We can similarly use this to isolate $\mathbf{R}$ by making assumptions about $\mathbf{B}$ or by using additional information in the analysis increments e.g. Desroziers et al. 2005 method.

An example: Satellite radiances from IASI

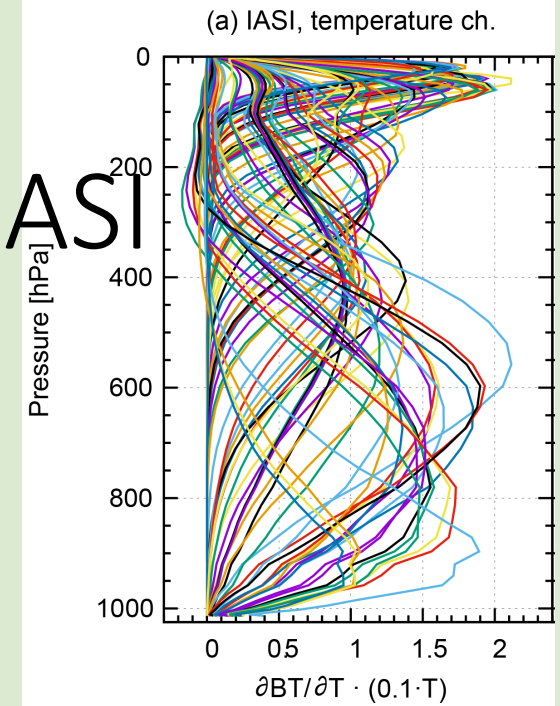
Satellite measurements of top of the atmosphere radiances for a given wavelength are sensitive to the temperature and gases (e.g. water vapour, ozone and carbon dioxide) throughout the atmospheric column as well as surface properties.

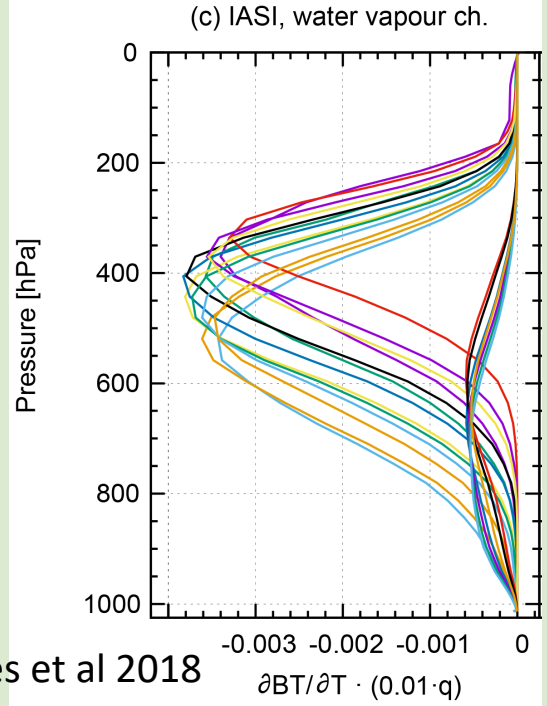
The mapping between the model variables and observed radiances can be achieved for each model column using fast radiative transfer models such as RTTOV. Right figure shows the Jacobian of the observation operator $\frac{\partial h(x)}{\partial x}$ for temperature and humidity for 123 IASI channels.

Left: For many instruments inter-channel error correlations are now represented in the R matrices.



Weston et al. 2013





Andrey-Andrés et al 2018

The forecast model

As we have already seen the forecast model is crucial to allowing us to cycle the assimilation system and produce our background estimate.

It may also be used to evolve the background error covariances either implicitly (see 4DVar) or explicitly (see EnKF).

If the observations assimilated are at different times it may also be incorporated into the observation operator.

Most geophysical models are non-linear.

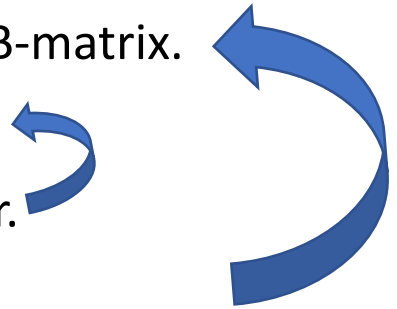
The non-linearity in the forecast models means that if the time between the analysis and the next background is too large then the assumption that the background errors are Gaussian may no longer hold.

Therefore, for methods that rely on the Gaussian assumption we need to be mindful of how frequently we perform the assimilation so that we remain in a quasi-linear regime.

Summary

All assimilation methods need the same basic ingredients to implement them:

- A background (first guess) and a representation of its error statistics, usually via the B-matrix.
- Observations and a representation of their error statistics, usually via the R-matrix.
- A way of comparing the model variables to those observed, the observation operator.
- A forecast model to allow for cycling of the assimilation.



Defining the error statistics and the observation operator can be incredibly difficult. Poor estimates of any of these will mean the observation cannot be used properly and information will be lost.

To make DA feasible we have made a lot of assumptions. In some situations, these may be quite limiting but can be overcome with extra levels of complexity e.g. Correction of observation and model biases, allowance for non-Gaussian distributions, estimation of key model parameters.

References

- Andrey-Andrés et al 2018: A simulated observation database to assess the impact of the IASI-NG hyperspectral infrared sounder, Atmos. Meas. Tech.
- Bannister 2008a: A review of forecast error covariance statistics in atmospheric variational data assimilation. I: Characteristics and measurements of forecast error covariances. QJRMS.
- Bannister 2008b: A review of forecast error covariance statistics in atmospheric variational data assimilation. II: Modelling the forecast error covariance statistics. QJRMS.
- Desroziers et al. 2005: Diagnosis of observation, background and analysis-error statistics
- in observation space. QJRMS.
- Fowler et al. 2018: On the interaction of observation and prior error correlations in data
- Assimilation. QJRMS.
- Hollingworth and Lönnberg, 1986: The statistical structure of shortrange forecast errors as determined from radiosonde data. Part I: The wind field. Tellus
- Parrish and Derber (1992)): The National Meteorological Center's spectral statistical interpolation analysis system. Mon. Weather Rev.
- Polavarapu et al. (2005): Data assimilation with the Canadian middle atmosphere model. Atmos.–Ocean
- Rainwater et al. 2015:The benefits of correlated observation errors for small scales. QJRMS
- Weston et al. 2013: Accounting for correlated error in the assimilation of high-resolution sounder data. QJRMS.