Basic concepts of data assimilation

Numerical Weather Prediction

NWP is an initial-boundary value problem:

- given an estimate of the present state of the atmosphere (initial conditions), and appropriate surface and lateral boundary conditions, the model simulates (forecasts) the atmospheric evolution.

The more accurate the estimate of the initial conditions, the better the quality of the forecasts.

Data Assimilation

The process of **combining *observations and short-range forecasts*** to obtain an initial condition for NWP **is called (atmospheric) data assimilation.**

The **purpose of data assimilation** is to determine as accurately as possible the state of the atmospheric flow by using all available information.

Objective Analysis – bringing observational data onto regular model grids

Early objective analysis methods – based on function fitting

Alternative definitions:

An analysis is the production of an accurate image of the true state of the atmosphere at a given time, represented by a collection of numbers, usually on regular model grids.

Objective analysis is an automated procedure for performing such analysis – versus subjective, hand analysis.

Data assimilation is an analysis technique in which the observed information is accumulated into the model state by taking advantage of consistency constraints with laws of time evolution and physical properties.

Explanation of Data Assimilation according by Andrew Lorenc (1995)

There are insufficient observations at any one time to determine the state of the atmosphere. So if we want a detailed complete picture, we need additional information. This is available as knowledge of the behavior and probable structure of the atmosphere. For instance, the knowledge of the typical structure of a frontal depression enables a human to draw an "analysis" of the atmospheric state, based on scattered observations.

To advance beyond this subjective approach, the behavior of the atmosphere is embodied in a computer model. In particular, knowledge of the evolution with time is embodied in a forecast model. This enables us to use observations distributed in time. The model also provides a consistent means of representing the atmosphere.

Assimilation is the process of finding the model representation of the atmosphere which is most consistent with the observations.

First Guess or Analysis Background and Analysis Cycle

Observations were often insufficient to determine all the unknowns in the initial condition – the problem is *underdetermined*. To solve this problem, **first guess** were introduced, first by Bergthorsson and Doos (1955).

The **first guess or background field** is our best estimate of the state of the atmosphere prior to the use of the observations.

The use of short-range forecasts as a first guess has been universally adopted in operational systems into what is called an "**analysis cycle**". Initially climatology, or a combination of climatology and a short forecast were used as a first guess.

In a well-behaved analysis/assimilation system, we expect that the analysis cycles allow the information to be accumulated in time into the model state, and to propagate to all variables (assuming all variables are not observed by every observing platforms) of the model.

Further explanation of the need to use background by Lorenc (1995):

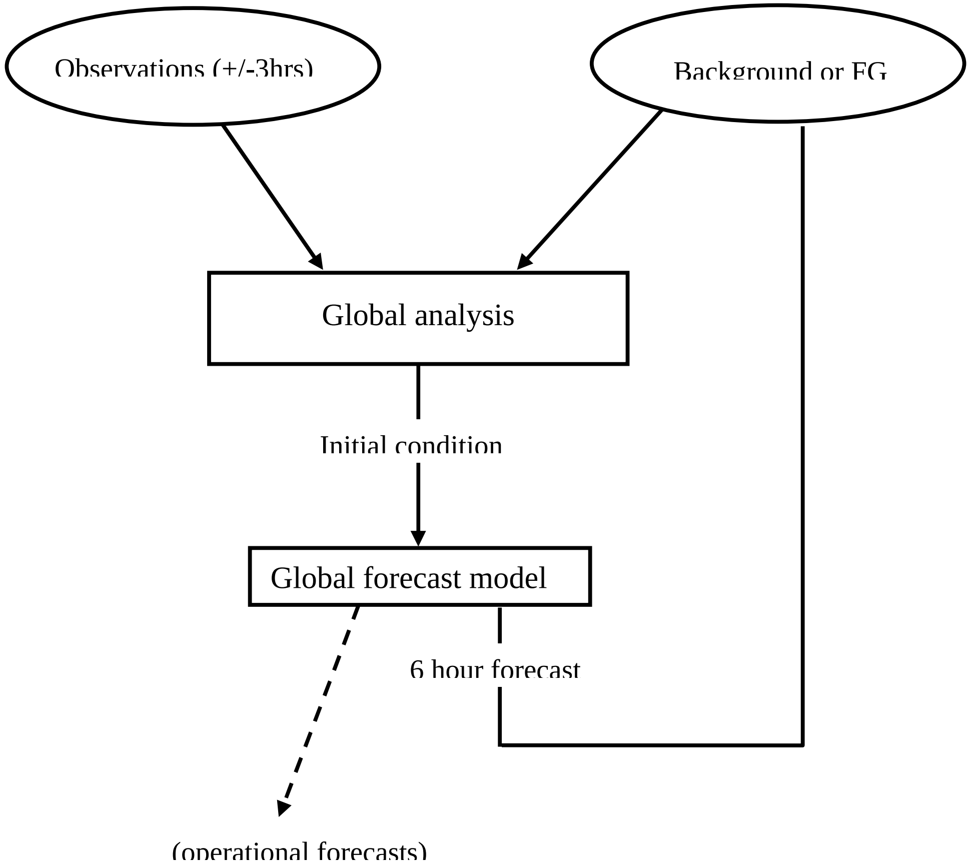
Usually, data assimilation proceeds sequentially in time. **The model organizes and propagates forward the information from previous observations.**

The information from new observations is used **to modify the model state, to be as consistent as possible with them (i.e., data) and the previous information**. It is the experience with operational assimilation for NWP that there is usually more information in the model state, from previous observations, than there is in a new batch of data at a single synoptic time. **Thus it is important to preserve this (i.e., information in the model data) in the assimilation process; it is not just a question of fitting the new data.**

Since all information has to be represented within the model, it is important that the model should be of sufficiently high resolution, with physically realistic detail, to represent the information observed.

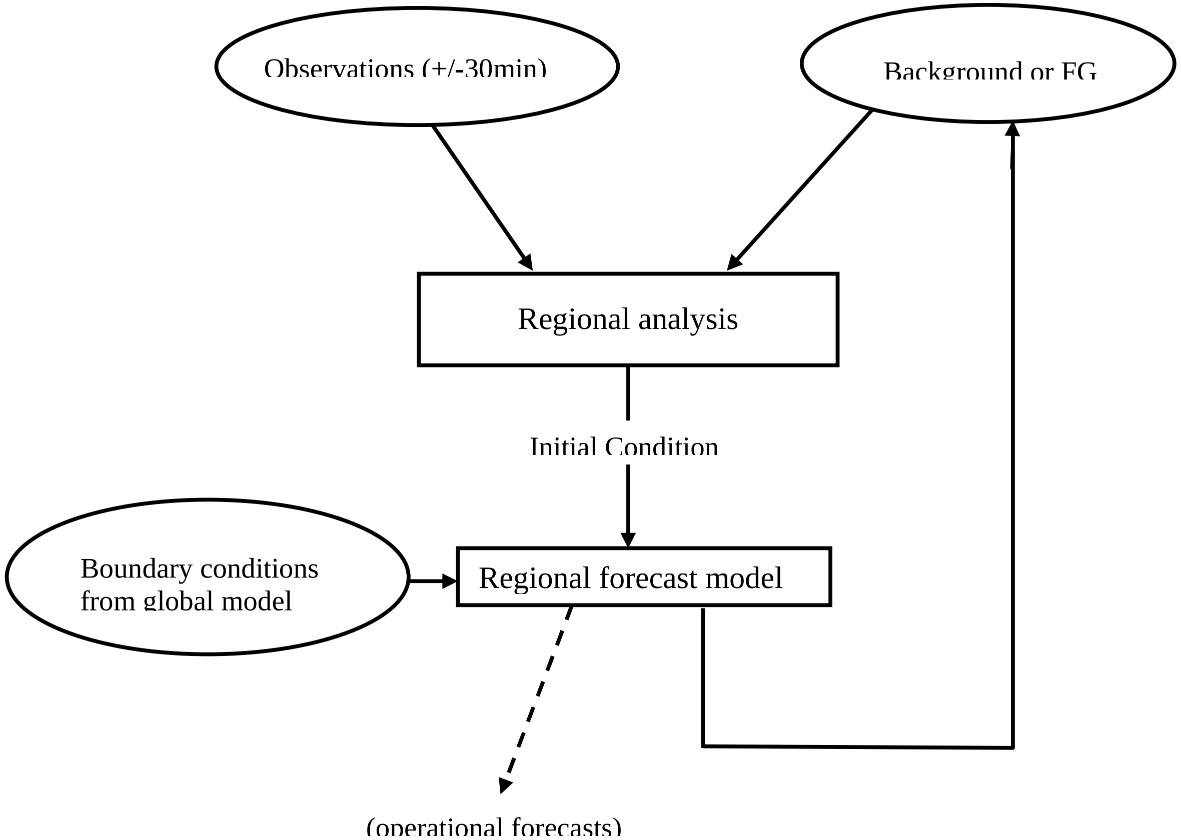
Some data assimilation methods, such as four-dimensional variational assimilation, are non-sequential. Still, the model state carries valuable information that should be preserved. Hence the need for assimilation cycles.

Typical Global 6-hour analysis cycle



Typical global 6-hr analysis cycle performed at 00, 06, 12, and 18 UTC at operational centers. The observations should be valid for the same time as the first guess. In the global analysis this has usually meant the rawinsondes launched mostly at the main observing times (00 and 12UTC), and satellite data are lumped into windows centered at the main observing times. The observations can be direct observations of variables used by the model, or indirect observations of geophysical parameters, such as radiances, that depend on the variables used in the model.

Typical Regional 1-hour analysis cycle



Typical regional analysis cycle. The main difference with the global cycle is that boundary conditions coming from global forecasts are an additional requirement for the regional forecasts

Two basic approaches to data assimilation

Sequential assimilation

only observations made in the past until the time of analysis are used, which is the case of real-time assimilation systems

Non-sequential, or retrospective assimilation

observation from the future can be used, for instance in a reanalysis exercise.

Intermittent or continuous data assimilation

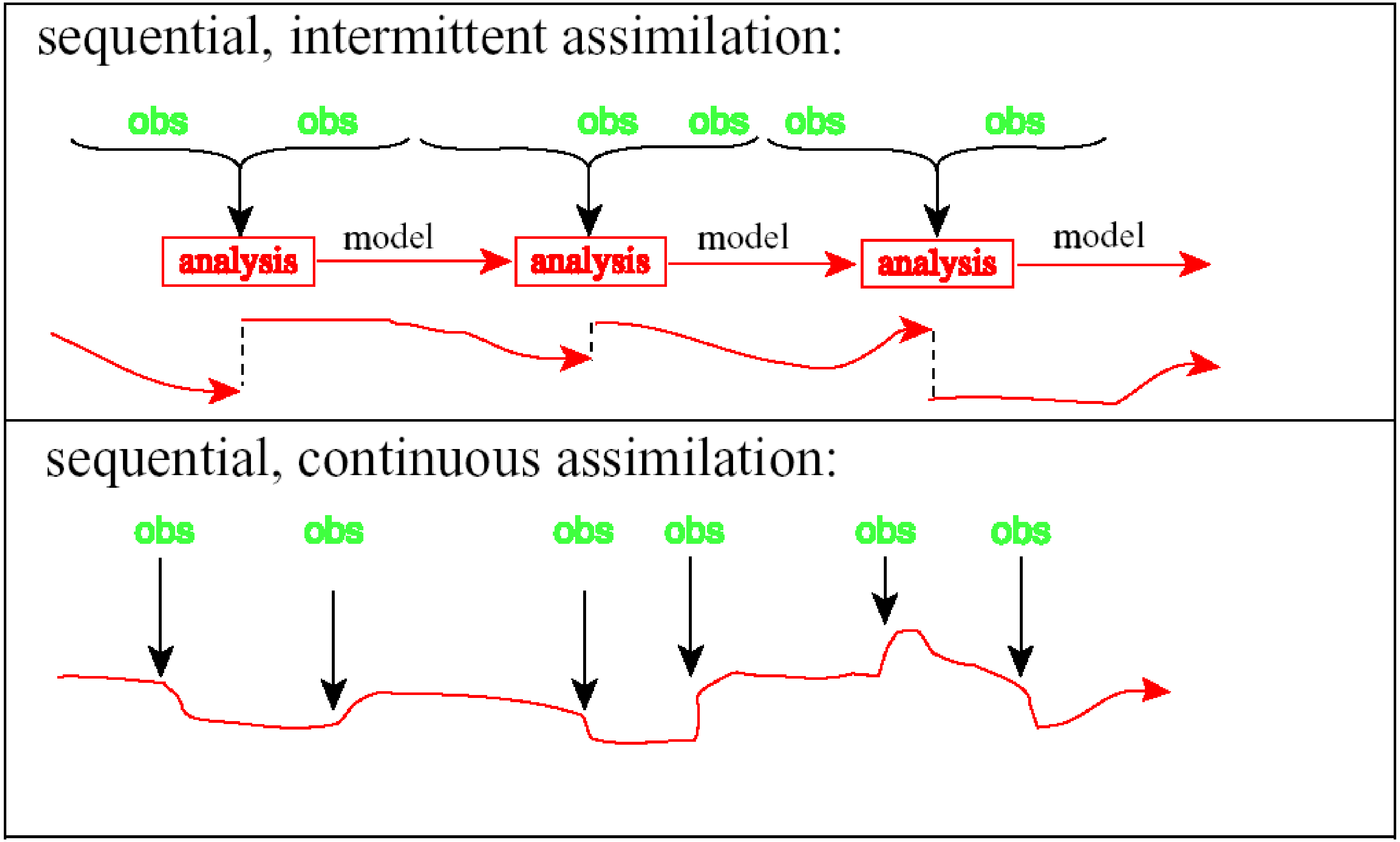
Assimilation methods can also be classified as **intermittent or continuous in time.**

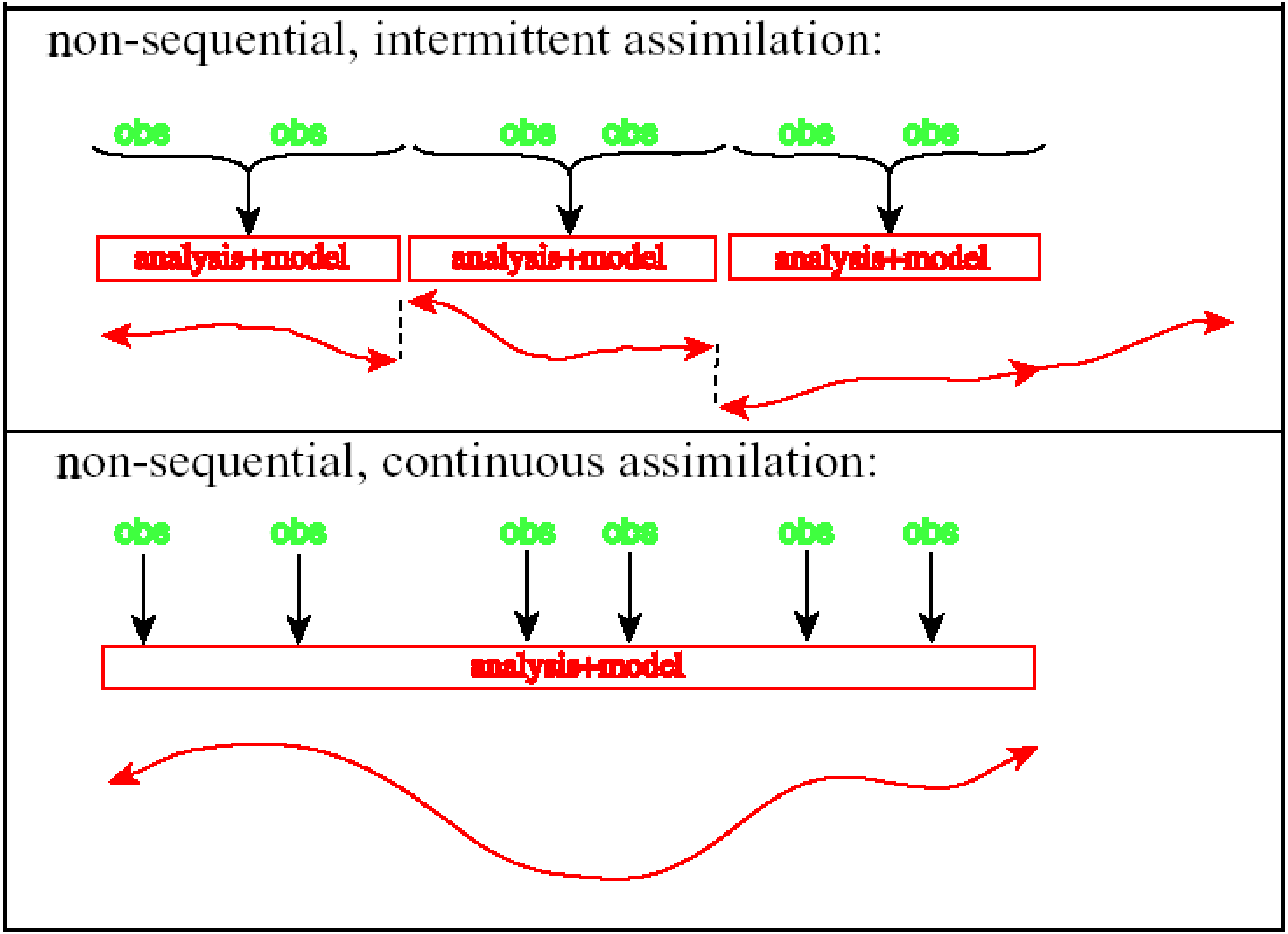
In an intermittent method, observations can be processed in small batches, which is usually technically convenient.

In a continuous method, observation batches over longer periods are considered, and the correction to the analysed state is smooth in time, which is physically more realistic.

Four basic strategies for data assimilation, as a function of time.

The way the time distribution of observations ("obs") is processed to produce a time sequence of assimilated states (the lower curve in each panel) can be sequential and/or continuous.





non-sequential, intermittent assimilation:

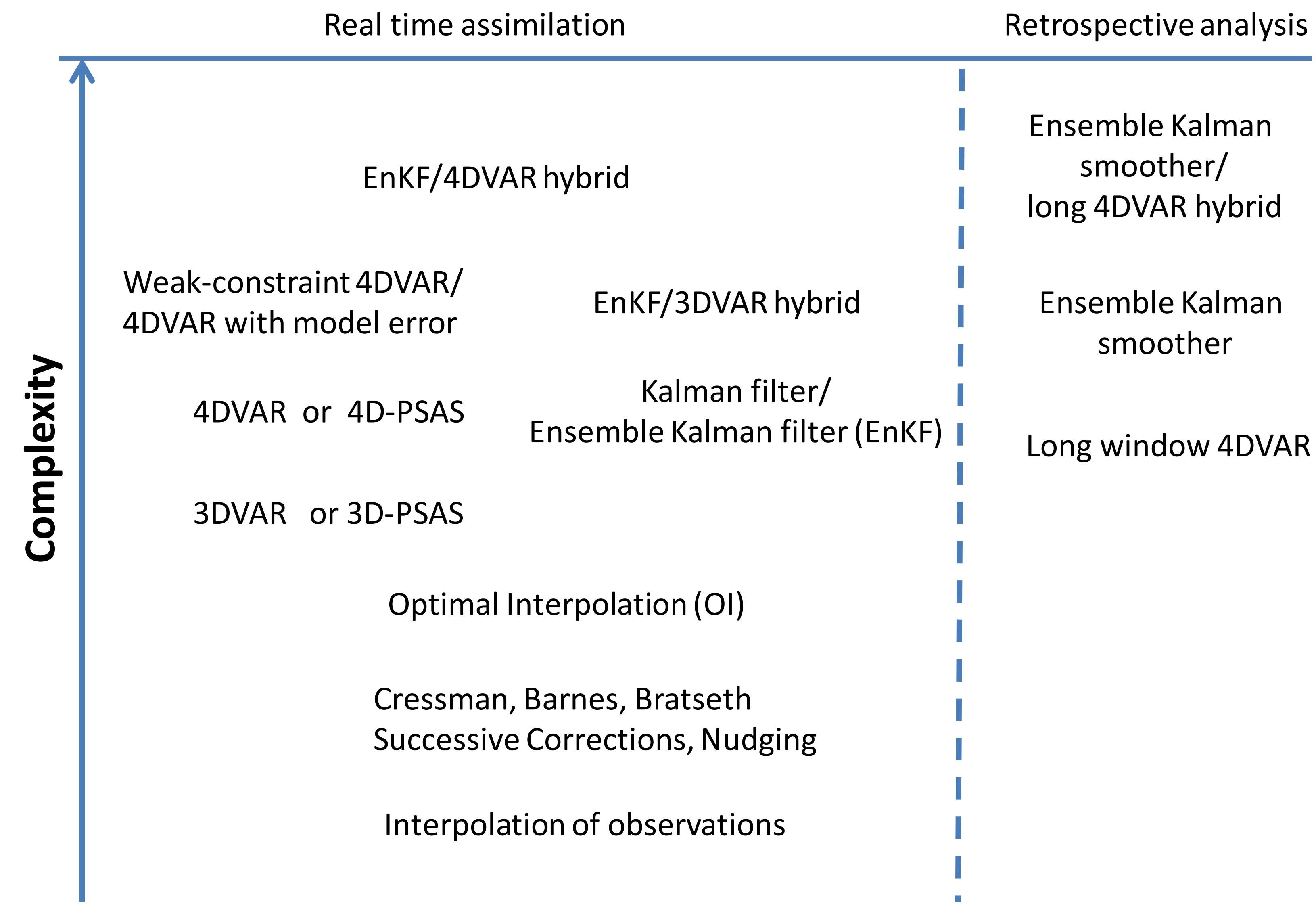
obs obs obs obs obs obs

analysis+model analysis+model analysis+model

non-sequential, continuous assimilation:

obs obs obs obs obs obs

analysis+model



Real time assimilation

Retrospective analysis

Complexity

EnKF/4DVAR hybrid

Ensemble Kalman smoother/ long 4DVAR hybrid

Weak-constraint 4DVAR/ 4DVAR with model error

EnKF/3DVAR hybrid

Ensemble Kalman smoother

4DVAR or 4D-PSAS

Kalman filter/ Ensemble Kalman filter (EnKF)

Long window 4DVAR

3DVAR or 3D-PSAS

Optimal Interpolation (OI)

Cressman, Barnes, Bratseth Successive Corrections, Nudging

Interpolation of observations

A summarized history of the main data assimilation algorithms used in meteorology and oceanography, roughly classified according to their complexity (and cost) of implementation, and their applicability to real-time problems. Currently, the most commonly used for operational applications are OI, 3DVAR, 4DVAR and ensemble Kalman filter (EnKF).

Many assimilation techniques have been developed for meteorology and oceanography (see figure above). They differ in their numerical cost, their optimality, and in their suitability for real-time data assimilation. This course deals with topics from Interpolation through 4DVAR in the left column. The data assimilation part of this course starts from Optimal Interpolation.

In addition, we will discuss the ensemble Kalman filter method that has emerged in very recent years to be a competitive alternative to 4DVAR. The EnKF-VAR hrbrid approach will be briefly discussed.

References:

Bergthorsson, P. and B. Doos, 1955: Numerical weather map analysis. *Tellus*, **7**.

Lorenc, A. C., 1986: Analysis methods for numerical weather prediction. *Quart. J. Roy. Meteor. Soc.*, **112,** 1177-1194.

Mathematical Preparation

Reading material: Sections in Chapter 2 of “Optimal Control and Estimation” by Robert Stengel.

Section 2.1 Scalars, vectors and matrices
Section 2.2 Matrix Properties and Operations
Section 2.4 Random variables, sequences and processes

Review of Linear Matrix Algebra

Scalar – is a single value /quantity

Vector – an order set of scalars that can be treated as single, multidimensional quantity is called a vector.

The number of elements/scalars in the vector defines the *dimension* of the vector.

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ . \\ x_n \end{bmatrix}$$

It has a dimension or vector length of *n*. It is a column vector.

The transpose of vector x is

$$\mathbf{x}^T = \begin{bmatrix} x_1 & x_2 & . & . & x_n \end{bmatrix}$$

It is a row vector.

Matrix - *two-dimensional arrays of scalars.*

A matrix **A** of dimension $n \times p$ is a 2-dimensional array of real coefficients $(a_{ij})_{i=1...n,\, j=1...p}$ where *i* is the line/row index and *j* is the column index.

$$\mathbf{A} = (a_{ij}) = \begin{bmatrix} a_{11} & a_{12} & . & . & a_{1p} \\ a_{21} & a_{22} & . & . & a_{2p} \\ . & . & . & . & . \\ . & . & . & . & . \\ a_{n1} & a_{n2} & . & . & a_{np} \end{bmatrix}$$

When *n* = *p*, the matrix is a **square** matrix. A vector of length *n* is an $n \times 1$ matrix.

Diagonal. The diagonal elements of a squared $n \times n$ matrix **A** contain *n* coefficients $(a_{ii})_{i=1.....n}$.

A matrix is called diagonal if all of its non-diagonal coefficients are zero.

Transpose. The transpose of an $n \times p$ matrix **A** is a $p \times n$ matrix denoted $\mathbf{A}^{\mathbf{T}}$ with the coefficients defined by $(a_{ij})^T = (a_{ji})$, i.e., the coefficients a_{ij} and a_{ji} are swapped.

$$\mathbf{A}^T = (a_{ji}) = \begin{bmatrix} a_{11} & a_{21} & . & . & a_{n1} \\ a_{12} & a_{22} & . & . & a_{n2} \\ . & . & . & . & . \\ . & . & . & . & . \\ a_{1p} & a_{2p} & . & . & a_{np} \end{bmatrix}$$

Symmetry. A square matrix is symmetric if it is equal to its transpose, i.e., $\mathbf{A} = \mathbf{A}^{\mathbf{T}}$. In this case, $a_{ij} = a_{ji}$ for all *i* and *j*.

Scalar multiplication. An $n \times p$ matrix **A** times a real scalar λ is defined as the $n \times p$ matrix λ **A** with coefficients (λa_{ij}).

Summation of matrices – matrices of identical dimension can be summed or differenced term by term:

$$\mathbf{C} = \mathbf{B} + \mathbf{A}$$

where **A**, **B**, and **C** have the same dimension. $c_{ij} = a_{ij} + b_{ij}$.

Product of matrices – the product between an $n \times p$ matrix **A** and a $p \times q$ matrix **B** is defined as the $n \times q$ matrix

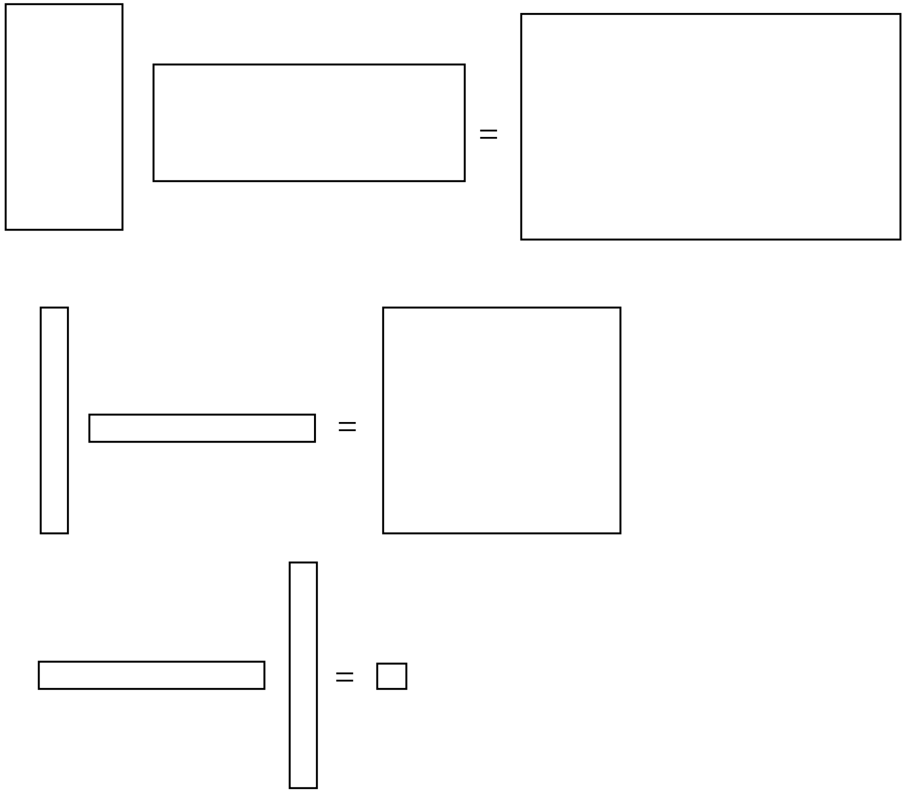
$$\mathbf{C} = \mathbf{A}\ \mathbf{B}$$

where

$$c_{ij} = \sum_{k=1}^{p} a_{ik} b_{kj} \; .$$

Note that $c_{ij} \neq a_{ij} b_{ij}$.

In the following, the rectangles represent the shapes of matrices:



If **x** and **y** are column vector of length *n*, $\mathbf{x}^{\mathbf{T}}\,\mathbf{y}$ is a scalar, while $\mathbf{x}\,\mathbf{y}^{\mathbf{T}}$ is an $n \times n$ square matrix!

The product is not defined if the number of columns in **A** is not the same as the number of rows in **B**.

Matrix multiplication is not commutative, i.e., **A B** ≠ **B A** except in special cases.

Identity matrix. An identity matrix **I** is a diagonal matrix whose elements/coefficients are all 1. Multiplication by an identity matrix has no effect.

Matrix inverse. A square $n \times n$ matrix is called invertible if there exists an $n \times n$ matrix denoted $\mathbf{A}^{\mathbf{-1}}$ and called inverse of **A**, such that $\mathbf{A}\ \mathbf{A}^{\mathbf{-1}} = \mathbf{A}^{\mathbf{-1}}\ \mathbf{A} = \mathbf{I}$**.**

For linear system of equations, **A x = y**, where **A** is the coefficient matrix, if the inverse $\mathbf{A}^{\mathbf{-1}}$ is found, the solution **x** is simply $\mathbf{x} = \mathbf{A}^{\mathbf{-1}}\ \mathbf{y}$, therefore the process of finding solution **x** is an inversion process.

Trace. The trace of a square $n \times n$ matrix **A** is defined as the scalar Tr(**A**) = $\sum_{i=1}^{n} a_{ii}$ which is the sum of diagonal coefficients.

Useful properties.

(**A**, **B**, **C** are assumed to be such that the operations below have a meaning)

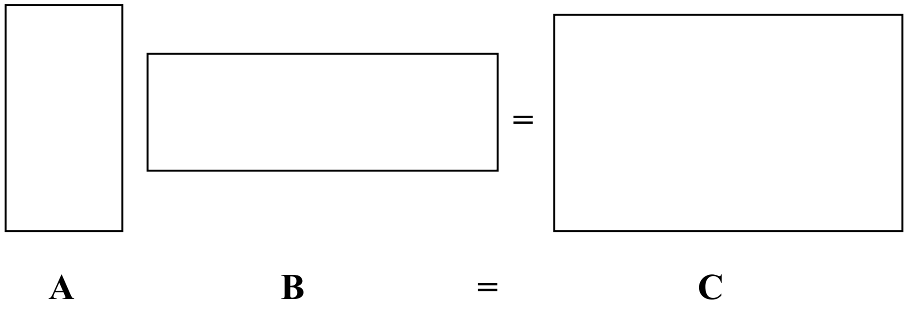
The transposition is linear: $(\mathbf{A} + \lambda\mathbf{B})^T = \mathbf{A}^T + \lambda\mathbf{B}^T$

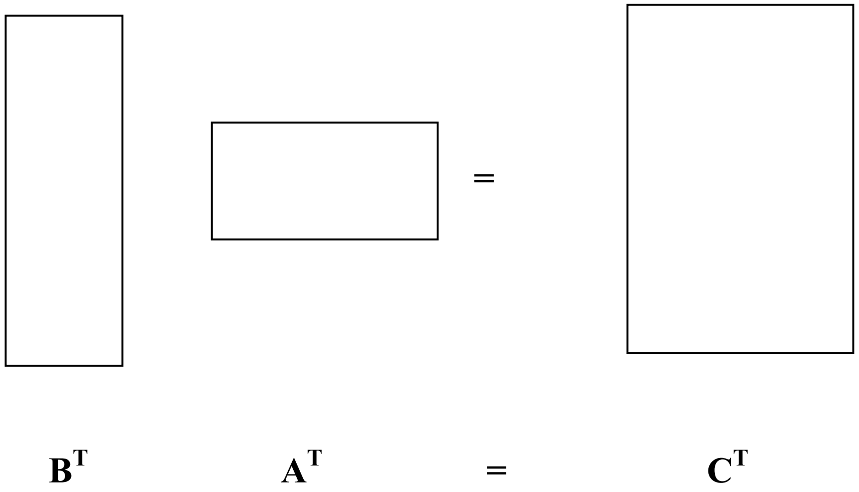
Transpose of a product: $(\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T$.

Let’s see if this makes sense:

Given an $n \times p$ matrix **A** and a $p \times q$ matrix **B**, the product **A B** is an $n \times q$ matrix and the dimension of $(\mathbf{A\ B})^{\mathbf{T}}$ is $q \times n$.

$\mathbf{A}^{\mathbf{T}}$ is a $p \times n$ matrix and $\mathbf{B}^{\mathbf{T}}$ is a $q \times p$ matrix, product $\mathbf{A}^{\mathbf{T}}\ \mathbf{B}^{\mathbf{T}}$ does not exist because the second dimension of $\mathbf{A}^{\mathbf{T}}$ does not equal to the first dimension of $\mathbf{B}^{\mathbf{T}}$ (generally). $\mathbf{B}^{\mathbf{T}}\ \mathbf{A}^{\mathbf{T}}$ exists because we have a $q \times p$ matrix multiplying a $p \times n$ matrix, and the result is a $q \times n$ matrix.




=
$\mathbf{B}^{\mathbf{T}}$
$\mathbf{A}^{\mathbf{T}}$
=
$\mathbf{C}^{\mathbf{T}}$

Inverse of a product: $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$

Inverse of a transpose: $(\mathbf{A}^T)^{-1} = (\mathbf{A}^{-1})^T$

Associativity of the product: **(A B) C = A (B C)**

Diagonal matrices: their products and inverses are diagonal, with coefficients given respectively by the products and inverses of the diagonals of the operands.

$$\begin{bmatrix} a_{11} & 0 & 0 & 0 \\ 0 & a_{22} & 0 & 0 \\ 0 & 0 & . & 0 \\ 0 & 0 & 0 & a_{nn} \end{bmatrix}\begin{bmatrix} b_{11} & 0 & 0 & 0 \\ 0 & b_{22} & 0 & 0 \\ 0 & 0 & . & 0 \\ 0 & 0 & 0 & b_{nn} \end{bmatrix} = \begin{bmatrix} a_{11}b_{11} & 0 & 0 & 0 \\ 0 & a_{22}b_{22} & 0 & 0 \\ 0 & 0 & . & 0 \\ 0 & 0 & 0 & a_{nn}b_{nn} \end{bmatrix}$$

$$\begin{bmatrix} a_{11} & 0 & 0 & 0 \\ 0 & a_{22} & 0 & 0 \\ 0 & 0 & . & 0 \\ 0 & 0 & 0 & a_{nn} \end{bmatrix}^{-1} = \begin{bmatrix} a_{11}^{-1} & 0 & 0 & 0 \\ 0 & a_{22}^{-1} & 0 & 0 \\ 0 & 0 & . & 0 \\ 0 & 0 & 0 & a_{nn}^{-1} \end{bmatrix} \text{ because } \begin{bmatrix} a_{11} & 0 & 0 & 0 \\ 0 & a_{22} & 0 & 0 \\ 0 & 0 & . & 0 \\ 0 & 0 & 0 & a_{nn} \end{bmatrix}\begin{bmatrix} a_{11}^{-1} & 0 & 0 & 0 \\ 0 & a_{22}^{-1} & 0 & 0 \\ 0 & 0 & . & 0 \\ 0 & 0 & 0 & a_{nn}^{-1} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & . & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Symmetric matrices: the symmetry is conserved by scalar multiplication, summation and inversion, but not by the product (in general).

The trace is linear: Tr (**A** +λ**B**) = Tr (**A**) + λTr(**B**)

Trace of a transpose: Tr($\mathbf{A}^{\mathbf{T}}$) = Tr(**A**)

Trace of a product: Tr(**AB)** = Tr(**BA**) = $\sum_{i,j} a_{ij} b_{ij}$.

Trace and basis change: Tr(**B**$^{-1}$**AB)** = Tr(**A**), i.e. the trace is an intrinsic property of the linear application represented by **A**.

Positive definite matrices. A symmetric matrix is defined to be positive definite if, for any vector **x**, the scalar $\mathbf{x}^{\mathbf{T}}\mathbf{A}\mathbf{x}$ > 0 unless **x**=0. Positive definite matrices have real positive eigenvalues, and their positive definiteness is conserved through inversion.