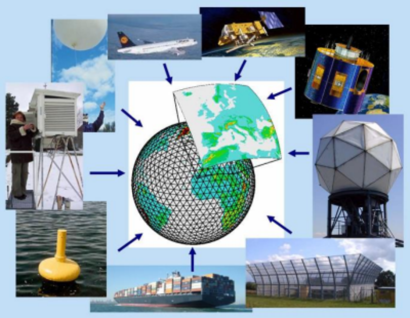
Numerical Weather Prediction: Data assimilation



Steven Cavallo

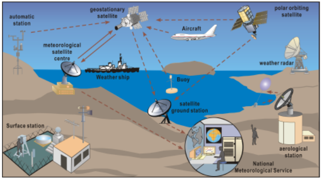
Data assimilation

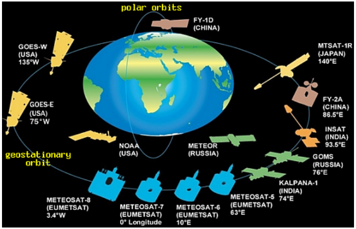
Data assimilation (DA) is the process estimating the **true state** of a system given **observations** of the system and a **background** estimate.

- Observations are not evenly spaced:
  - ➔ MUCH greater number of observations at surface than aloft.
  - ➔ Fewer observations over oceans.
  - ➔ Observations, themselves, have error (e.g. instrument error).
- In order to predict the future, the current state MUST be known.
  - ➔ Future state $=$ Current state + change in current state
- Idea is that better initial conditions (ICs) $\Rightarrow$ better forecast:
  - ➔ Forecast error = Model error + IC error
- DA helps constrain the model to better fit observations.
- DA is a statistical combination of observations and short-term model forecasts.

Data assimilation

- Observations come from a variety of places, including surface stations, satellites, radiosondes, commercial aircraft, buoys, radar, mesonet sites, ships, and more.
- Observations have varying degrees of instrument error, as well as processing error (e.g. satellite and radar data).
- Once observations are obtained, they are checked through a **quality control** process. “Bad” observations are filtered out statistically by comparing the observations value with the model’s first guess, and using the known error characteristics of that particular observation.





Data assimilation

The data assimilation problem can be thought of as determining the **probability density function (PDF)** of the current state given all current and past observations:

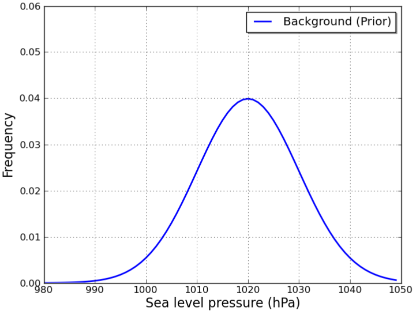
$$\underbrace{P\left(X_t^t \mid Y_t\right)}_{\text{Analysis (Posterior)}} \propto \underbrace{P\left(Y_t \mid X_t^t\right)}_{\text{Observations}} * \underbrace{P\left(X_t^t \mid Y_{t-1}\right)}_{\text{Background (Prior)}} \tag{1}$$

$$\begin{aligned} X_t^t &: \quad \text{Current state} \\ Y_t &: \quad \text{Current and past observations} \\ Y_{t-1} &: \quad \text{Past observations} \end{aligned}$$

The **background**, or **prior**, is a first guess of the analysis. Usually, it is 6-hour model forecasts.

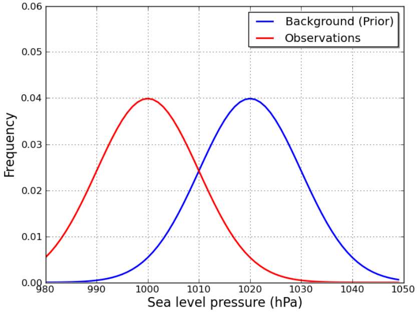
Data assimilation

$$\underbrace{P\left(X_t^t \mid Y_t\right)}_{\text{Analysis (Posterior)}} = \underbrace{P\left(Y_t \mid X_t^t\right)}_{\text{Observations}} * \underbrace{P\left(X_t^t \mid Y_{t-1}\right)}_{\text{Background (Prior)}}$$



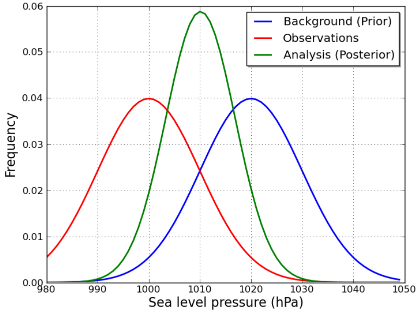
Data assimilation

$$\underbrace{P\left(X_t^t \mid Y_t\right)}_{\text{Analysis (Posterior)}} = \underbrace{P\left(Y_t \mid X_t^t\right)}_{\text{Observations}} * \underbrace{P\left(X_t^t \mid Y_{t-1}\right)}_{\text{Background (Prior)}}$$



Data assimilation

$$\underbrace{P\left(X_t^t \mid Y_t\right)}_{\text{Analysis (Posterior)}} = \underbrace{P\left(Y_t \mid X_t^t\right)}_{\text{Observations}} * \underbrace{P\left(X_t^t \mid Y_{t-1}\right)}_{\text{Background (Prior)}}$$



Data assimilation

Notation:

x^a	:	Model analysis
x^b	:	Model background (short-term forecasts)
x^o	:	Observations
x^t	:	"True state"
σ_b^2	:	Background error variance
σ_o^2	:	Observation error variance

A model analysis is made using *Linear analysis*, or a linear combination of the observations and the model's first guess of the atmospheric state:

$$x^a = a_1 x^b + a_2 x^o. \tag{2}$$

If we assume that there is no mean bias in the observations or background (but that we know the variance of the background and observational error), then the weights a_1 and a_2 can be chosen in a way that minimizes the mean squared error of x^a:

$$a_1 = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_o^2};\, a_2 = \frac{\sigma_o^2}{\sigma_b^2 + \sigma_o^2}. \tag{3}$$

Data assimilation

Defining a weighting function as

$$W \equiv \frac{\sigma_b^2}{\sigma_b^2 + \sigma_o^2},$$

then

$$1 - W = \frac{\sigma_o^2}{\sigma_b^2 + \sigma_o^2}$$

so that the analysis equation (2) becomes

$$\boxed{x^a = x^b + W\left(x^o - x^b\right)}. \tag{4}$$

Some more terms:

$$\begin{aligned} x^a - x^b \quad &= \quad \text{Analysis increment} \\ x^o - x^b \quad &= \quad \text{Innovation (new information)} \end{aligned}$$

Data assimilation

W contains the **background error covariance**. Most data assimilation techniques today differ in how they treat this background error covariance.

1. Statistical interpolation (SI)
   - W prescribed by distance from observation.
   - No error information. Simply the interpolation of observations onto a grid.

Data assimilation

W contains the **background error covariance**. Most data assimilation techniques today differ in how they treat this background error covariance.

1. Statistical interpolation (SI)
   - W prescribed by distance from observation.
   - No error information. Simply the interpolation of observations onto a grid.
2. Optimum Interpolation (OI)
   - $W \equiv K = BH^T \left(R + HBH^T\right)^{-1}$.
   - B is the error covariance, but it is fixed.
   - Observation influence is limited to small region near the observations.

Data assimilation

W contains the **background error covariance**. Most data assimilation techniques today differ in how they treat this background error covariance.

1. Statistical interpolation (SI)
   - W prescribed by distance from observation.
   - No error information. Simply the interpolation of observations onto a grid.
2. Optimum Interpolation (OI)
   - $W \equiv K = BH^T \left(R + HBH^T\right)^{-1}$.
   - B is the error covariance, but it is fixed.
   - Observation influence is limited to small region near the observations.
3. 3DVAR
   - B is fixed, so observation impact is isotropic around observation.
   - It does not directly solve matrices. This makes it computationally easy and efficient.

Data assimilation

W contains the **background error covariance**. Most data assimilation techniques today differ in how they treat this background error covariance.

1. Statistical interpolation (SI)
   - W prescribed by distance from observation.
   - No error information. Simply the interpolation of observations onto a grid.
2. Optimum Interpolation (OI)
   - $W \equiv K = BH^T \left(R + HBH^T\right)^{-1}$.
   - B is the error covariance, but it is fixed.
   - Observation influence is limited to small region near the observations.
3. 3DVAR
   - B is fixed, so observation impact is isotropic around observation.
   - It does not directly solve matrices. This makes it computationally easy and efficient.
4. EnKF
   - B is flow-dependent.
   - Ensemble method—everything is a matrix. This is computationally expensive.

Data assimilation

Using 3DVAR or 4DVAR, matrices are not solved. Instead, a cost function is defined to describe the distance between the observations, background, and 'true' state, and this cost function is minimized to produce a single analysis. 4DVAR differs from 3DVAR in that different times are taken into account.

Currently, ECMWF uses 4DVAR. GFS used 3DVAR until May 2012, when it uses a "hybrid" 3DVAR and EnKF.
The **Ensemble Kalman Filter (EnKF)** utilizes an **ensemble** of model forecasts.

$$\widetilde{x^a} = \widetilde{x^b} + K\left(\widetilde{x^o} - H(\widetilde{x^b})\right) \tag{5}$$

where the $\sim$ symbols denotes an array (or ensemble), $H(\widetilde{x^b})$ means it is the interpolation between the model grid and observation space, and K is called the **Kalman gain** matrix:

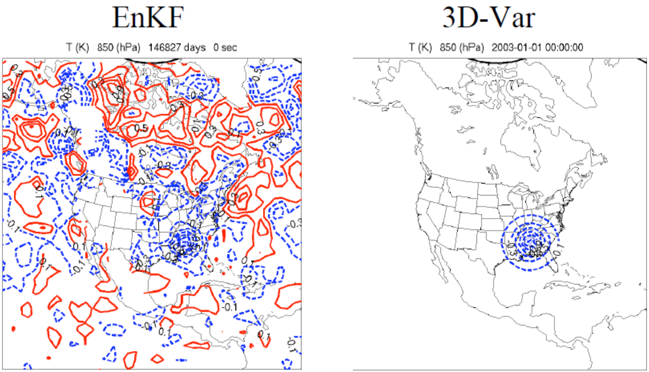
$$K = BH^T\left(R + HBH^T\right)^{-1}. \tag{6}$$

The background error covariance is B and the observations error covariance is R.

Data assimilation

With EnKF, the background error covariance matrix depends on the atmospheric state, since it is simply the model error ($B = \text{cov}(\epsilon^b, \epsilon^b)$) where $\epsilon^b = \widetilde{x^b} - \widetilde{x^{true}}$. In 3DVAR, B is usually a climatology that does not get updated.

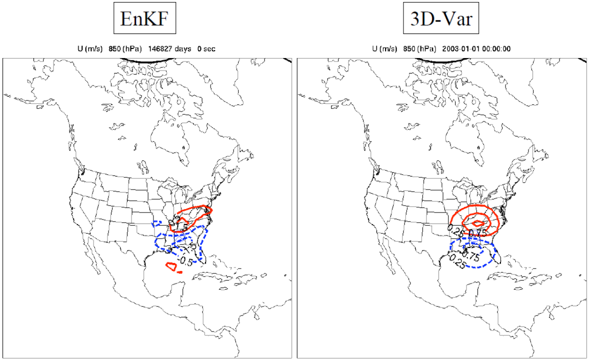
850 hPa temperature analysis increment



Data assimilation

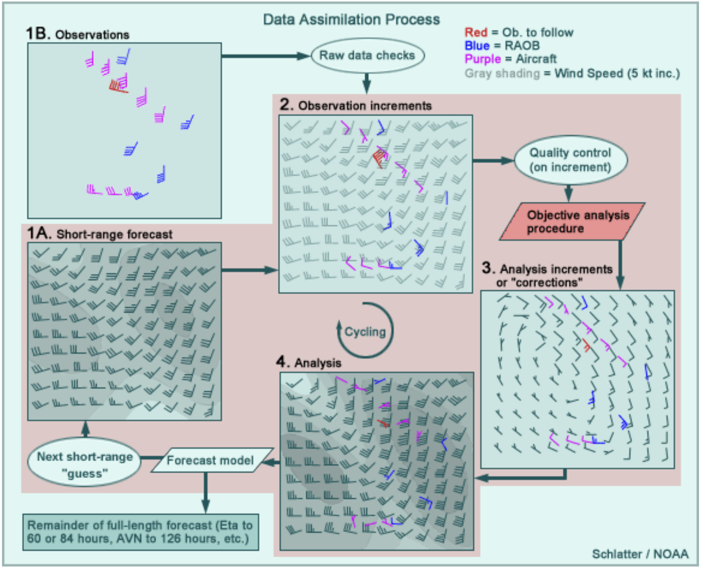
With EnKF, the background error covariance matrix depends on the atmospheric state, since it is simply the model error ($\mathrm{B} = \mathrm{cov}(\epsilon^b, \epsilon^b)$) where $\epsilon^b = \widetilde{x^b} - \widetilde{x^{true}}$. In 3DVAR, B is usually a climatology that does not get updated.

850 hPa U analysis increment



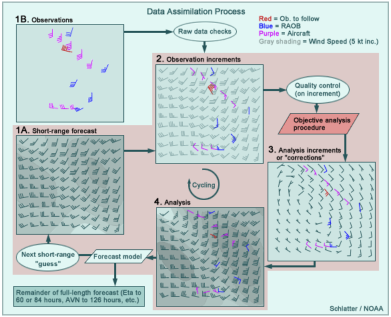
Data assimilation

Summary of the data assimilation process



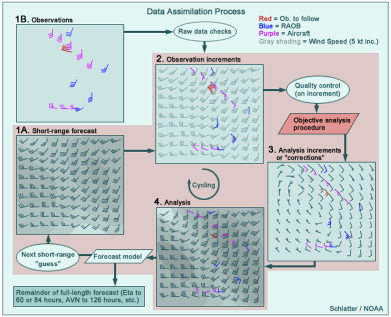
Data assimilation

1. Gather observations and make a short-term model forecast.



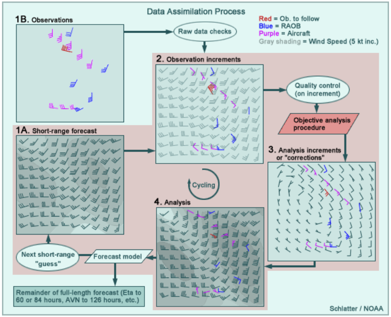
Data assimilation

1. Gather observations and make a short-term model forecast.
2. Compute observation increment. This is the difference between the observed data and the background data after the background data has been converted to observation space (via time and space interpolation). This must be done in order to perform quality control checks.



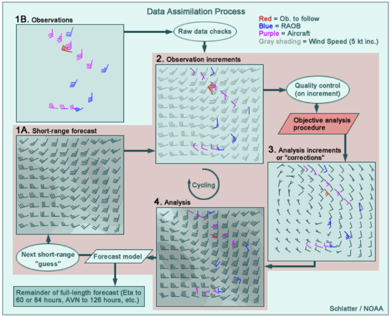
Data assimilation

1. Gather observations and make a short-term model forecast.
2. Compute observation increment. This is the difference between the observed data and the background data after the background data has been converted to observation space (via time and space interpolation). This must be done in order to perform quality control checks.
3. Merge observation increments to model grid and compute analysis increment $K\left(\widetilde{x^o} - H(\widetilde{x^b})\right)$.



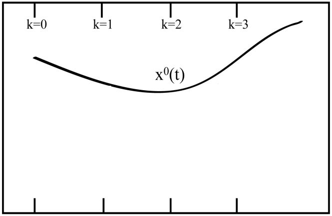
Data assimilation

1. Gather observations and make a short-term model forecast.
2. Compute observation increment. This is the difference between the observed data and the background data after the background data has been converted to observation space (via time and space interpolation). This must be done in order to perform quality control checks.
3. Merge observation increments to model grid and compute analysis increment $K\left(\widetilde{x^o} - H(\widetilde{x^b})\right)$.
4. Compute analysis.



Data assimilation

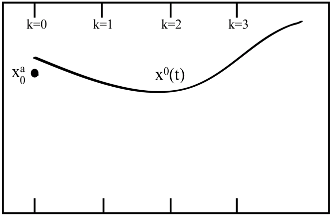
Schematic



$x^o(t)$: Observations of x at time t

Data assimilation

Schematic

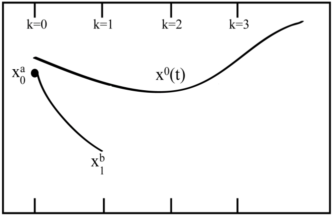


$x^o(t)$: Observations of x at time t

x_0^a: Analysis at time $k = 0$

Data assimilation

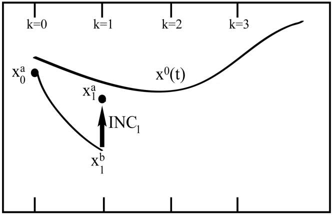
Schematic



$x^o(t)$:	Observations of x at time t
x_0^a:	Analysis at time $k=0$
x_1^b:	Background forecast at time $k=1$

Data assimilation

Schematic



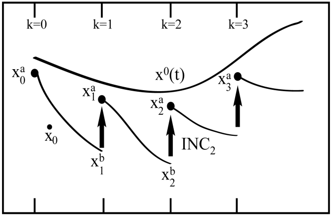
$x^o(t)$:	Observations of x at time t
x^a_0:	Analysis at time $k = 0$
x^b_1:	Background forecast at time $k = 1$

$$x^a_1 = x^b_1 + K\left(x^o_1 - H(x^b_1)\right)$$

$$INC_1 = x^a_1 - x^b_1$$

Data assimilation

Schematic



$x^o(t)$: Observations of x at time t

x_0^a: Analysis at time $k = 0$

x_1^b: Background forecast at time $k = 1$

$$x_1^a = x_1^b + K\left(x_1^o - H(x_1^b)\right)$$

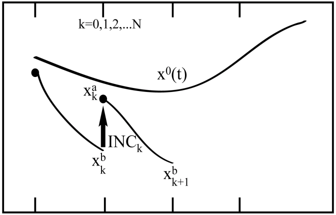
$$INC_1 = x_1^a - x_1^b$$

or more generally

$$INC_k = x_k^a - x_k^b$$

Data assimilation

Schematic



$x^o(t)$:	Observations of x at time t
x_0^a:	Analysis at time $k = 0$
x_1^b:	Background forecast at time $k = 1$

$$
\begin{aligned}
x_1^a &= x_1^b + K\left(x_1^o - H(x_1^b)\right) \\
INC_1 &= x_1^a - x_1^b
\end{aligned}
$$

or more generally

$$INC_k = x_k^a - x_k^b$$

Data assimilation

Example: EnKF data assimilation for tropical cyclone prediction

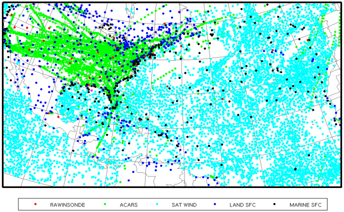
- WRF ARW v. 3.1, 36 km horizontal resolution, 96 ensemble members
- DART assimilation system, based on Ensemble Kalman Filter (EnKF)
- Assimilates surface and marine stations, rawindsondes, ACARS, satellite winds, TC position and minimum sea level pressure every 6 hours

Data assimilation

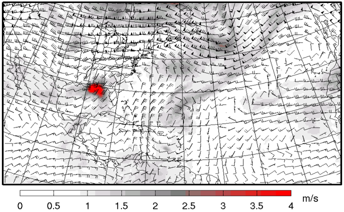
Example: EnKF data assimilation for tropical cyclone prediction

- WRF ARW v. 3.1, 36 km horizontal resolution, 96 ensemble members
- DART assimilation system, based on Ensemble Kalman Filter (EnKF)
- Assimilates surface and marine stations, rawindsondes, ACARS, satellite winds, TC position and minimum sea level pressure every 6 hours

Observations assimilated



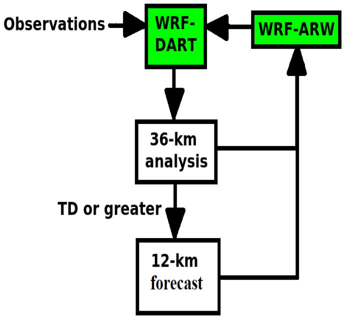
850-200 hPa wind



00 UTC 10 Nov. 2009

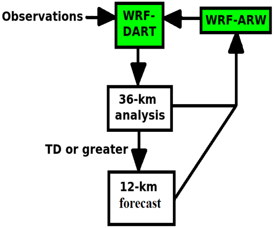
Data assimilation

- WRF ARW v. 3.1, 36 km horizontal resolution, 96 ensemble members
- DART assimilation system, Ensemble Kalman Filter (EnKF)
- Assimilates surface and marine stations, rawindsondes, ACARS, satellite winds, TC position and minimum sea level pressure every 6 hours



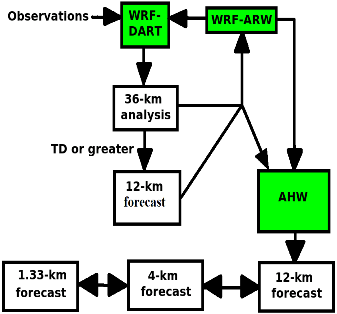
Data assimilation

- Cycled continuously from August 10, 2009 - November 10, 2009
- If NHC declares a tropical depression or stronger, a 12-km nest is created
- Initial condition for high resolution forecast from the ensemble member closest to observation



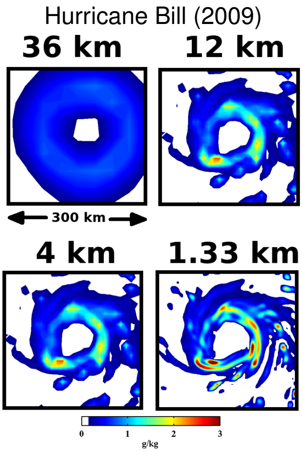
Advanced Hurricane WRF (AHW) forecasts

- Based on WRF v. 3.1, initial conditions from WRF-DART
- 12-km parent domain, Kain-Fritsch cumulus scheme
- 4-km, 1.33-km nests, no cumulus parameterization, following storm
- RRTM longwave, Dudhia shortwave, WSM-5 microphysics, YSU boundary layer
- 36 vertical levels, 1-D Ocean



Data assimilation

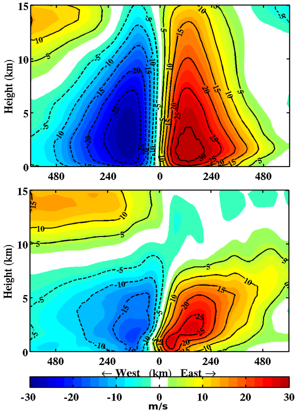
- Reduce forecast errors with:
  - ➔ High resolution forecasts
    - ➔ Resolve details of storms, such as eyewall structure, bands



Data assimilation

- Reduce forecast errors with:
  - ➔ High resolution forecasts
    - ➔ Resolve details of storms, such as eyewall structure, bands
- Improved initial conditions
  - ➔ Asymmetries, vertical tilt
  - ➔ Vortex not pre-defined, minimal model spin-up

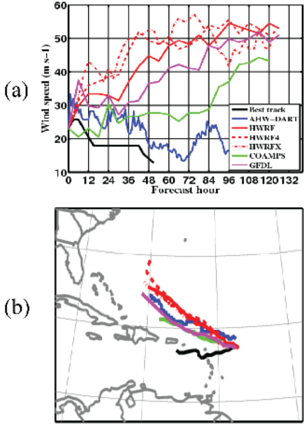
Tropical storm Erika (2009)



Data assimilation

- Reduce forecast errors with:
  - ➔ High resolution forecasts
    - ➔ Resolve details of storms, such as eyewall structure, bands
- Improved initial conditions
  - ➔ Asymmetries, vertical tilt
  - ➔ Vortex not pre-defined, minimal model spin-up

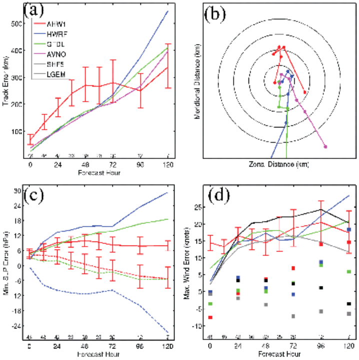
Tropical storm Erika (2009)



Data assimilation

AHW forecast verification

- AHW comparable to HWRF, GFDL, GFS, and others.
- Cyclone track error is large in short-term forecasts, but better at long-term forecasts.
- Intensity error smaller than HWRF or GFDL forecasts.



Data assimilation

Key aspects of data assimilation (DA):

- DA is how the weighting between observations and short-term forecasts (background) is performed to create an analysis.

Data assimilation

Key aspects of data assimilation (DA):

- DA is how the weighting between observations and short-term forecasts (background) is performed to create an analysis.
- The weighting depends on accurate knowledge of the error that is associated with the background forecasts (background error covariance) and observations (observation error covariance).

Data assimilation

Key aspects of data assimilation (DA):

- DA is how the weighting between observations and short-term forecasts (background) is performed to create an analysis.
- The weighting depends on accurate knowledge of the error that is associated with the background forecasts (background error covariance) and observations (observation error covariance).
- Top methods:
  - ➔ 3DVAR: Used for GFS until May 2012. Assumes constant error statistics (Fixed background error covariances).
  - ➔ 4DVAR: Currently used by ECMWF. Similar to 3DVAR, except observations from different times are incorporated.
  - ➔ EnKF: Uses an ensemble to provide flow-dependent background error covariances. Ensemble also provides **probabilistic** representation of the initial state and forecasts. However, this is computationally expensive.

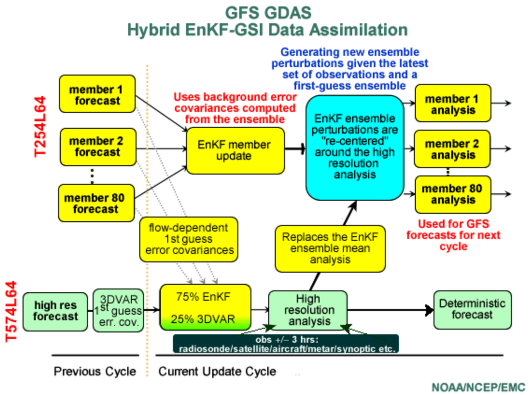
Data assimilation

Key aspects of data assimilation (DA):

- DA is how the weighting between observations and short-term forecasts (background) is performed to create an analysis.
- The weighting depends on accurate knowledge of the error that is associated with the background forecasts (background error covariance) and observations (observation error covariance).
- Top methods:
  - ➔ 3DVAR: Used for GFS until May 2012. Assumes constant error statistics (Fixed background error covariances).
  - ➔ 4DVAR: Currently used by ECMWF. Similar to 3DVAR, except observations from different times are incorporated.
  - ➔ EnKF: Uses an ensemble to provide flow-dependent background error covariances. Ensemble also provides **probabilistic** representation of the initial state and forecasts. However, this is computationally expensive.
- Hybrid EnKF: Currently used to create GFS analyses.

Data assimilation

The Hybrid EnKF uses EnKF to create an ensemble of short-term forecasts that provides flow-dependent covariances.

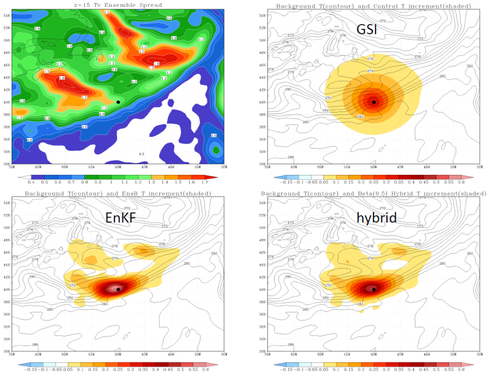


Data assimilation

The GFS data assimilation system (GDAS) Hybrid-EnKF upgrade was implemented in May 2012.

GFS forecasts have improved, as seen by the 500 hPa height anomaly correlation skill score and in tropical cyclone forecast tracks.

Single 850 hPa T observation:
Analysis increment

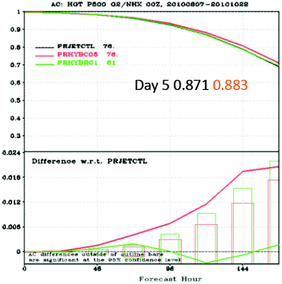


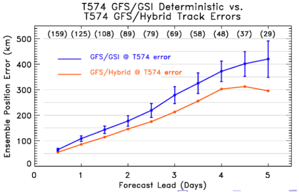
Data assimilation

The GFS data assimilation system (GDAS) Hybrid-EnKF upgrade was implemented in May 2012.

GFS forecasts have improved, as seen by the 500 hPa height anomaly correlation skill score and in tropical cyclone forecast tracks.

500 hPa anomaly correlation:
(Red = Hybrid, Black = Old GDAS)





Important points and questions for review

- What are 3 ways that model error can be introduced into a forecast?
- What is the analysis equation? What is an analysis increment and innovation?
- What is a background error covariance?
- What is the primary difference in how data assimilation systems differ?
- Until May 2012, GFS used the 3DVAR data assimilation method. However, EnKF has been shown to have lower analysis and forecast error. What are the differences between 3DVAR and EnKF? Why do you think a hybrid EnKF was implemented in May 2012 instead of a full EnKF?