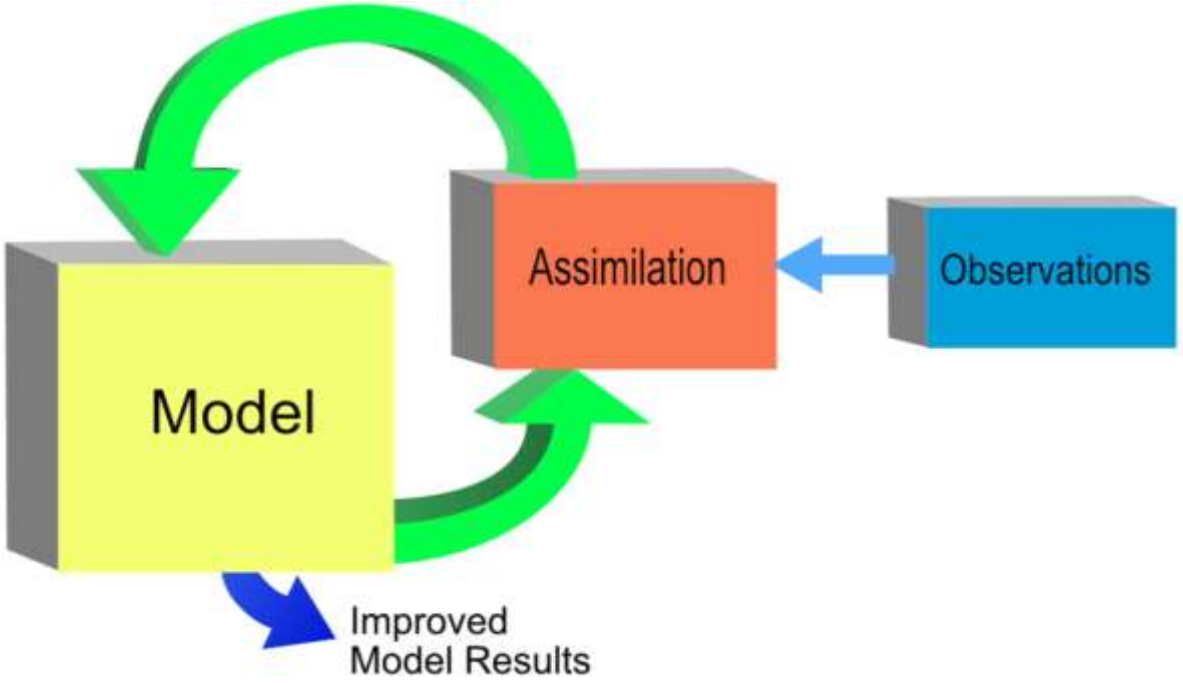
BASIC CONCEPTS OF OCEAN DATA ASSIMILATION



"Fundamentals of Ocean Modeling" during 27 September - 1 October 2021
organized by
International Training Centre for Operational Oceanography (ITCOocean),
ESSO-INCOIS, Hyderabad, India

ARYA PAUL
Scientist E

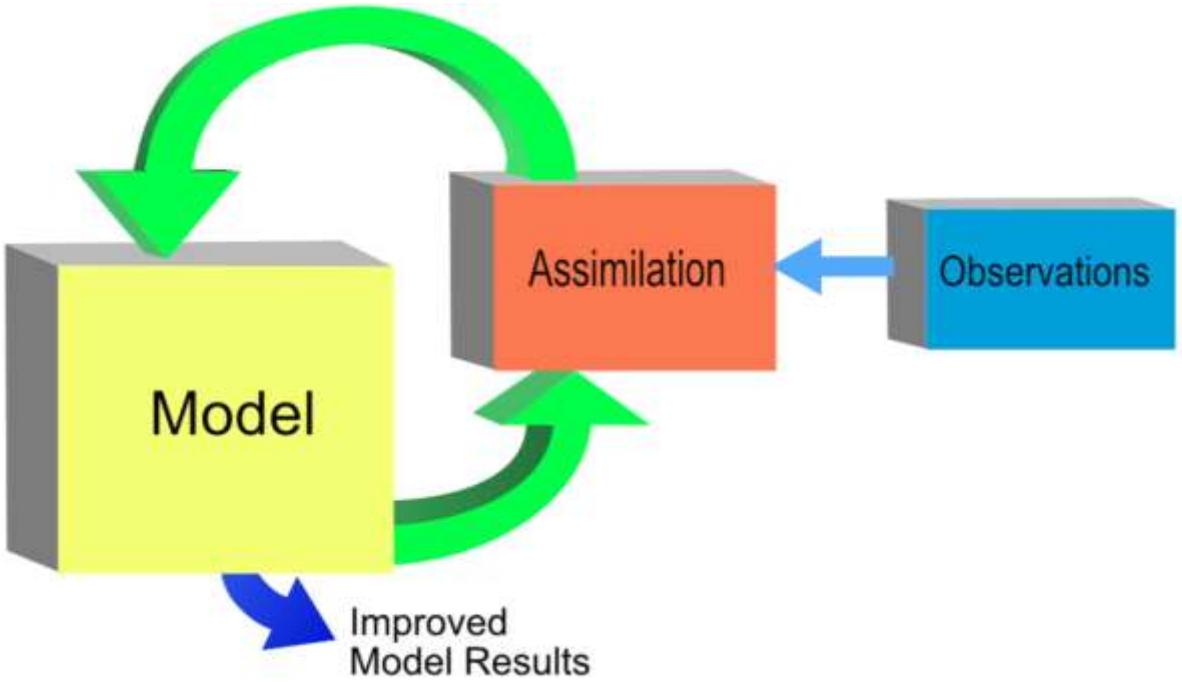
Gauss, 1809: Theoria Motus Corporum Coelestium -1823: Theoria combinationis Observationum erroribus minimis obnoxiae

But since our measurements and observations are nothing more than approximations to truth, the same must be true of all calculations resting upon them, and the highest aim of all computation made concerning concrete phenomena must be to approximate, as nearly as practicable to the truth.

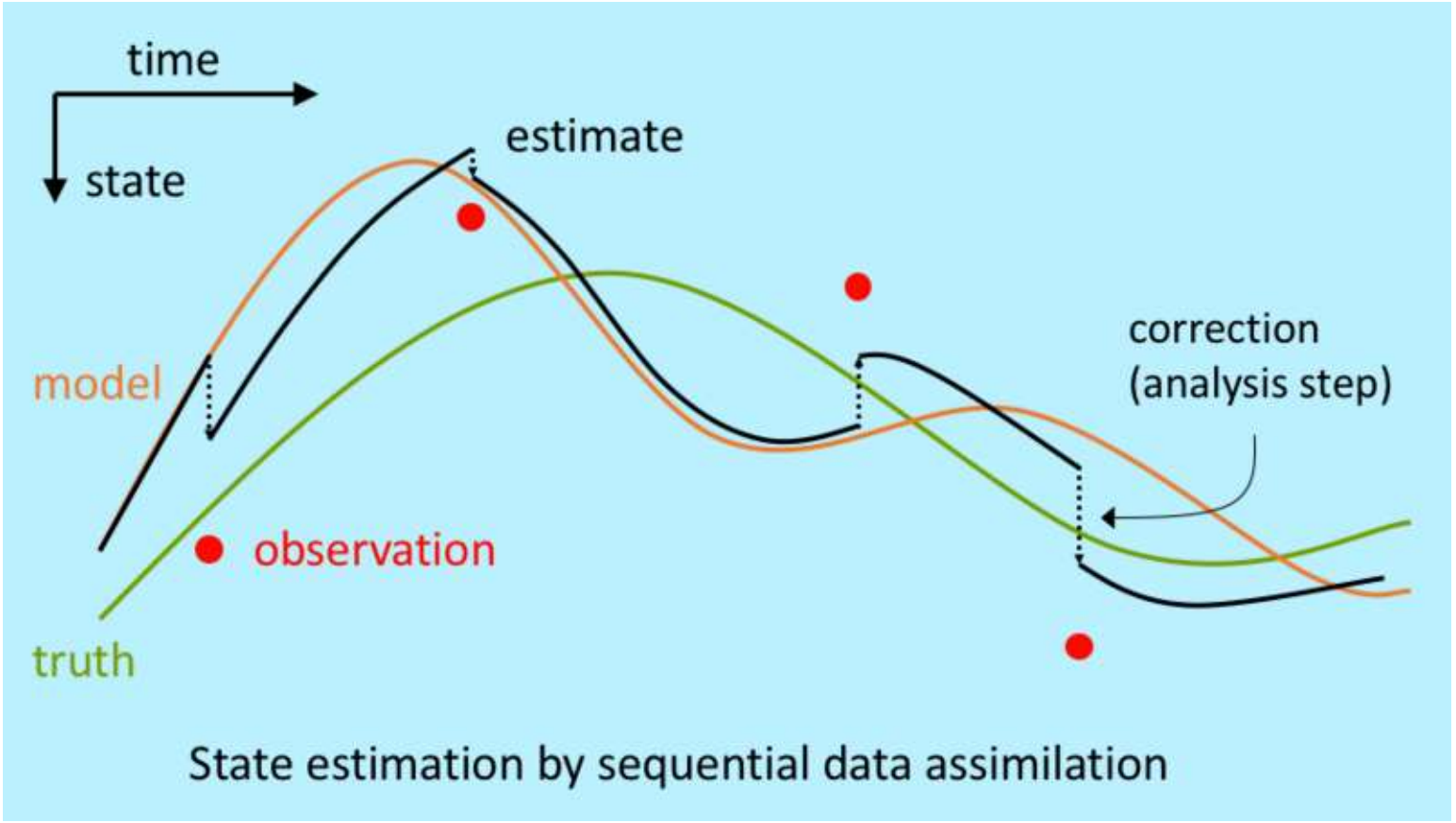
Summary of Gauss's idea

- Observations are approximate.
- Truth is not known.
- Model also has errors due to various reasons.

What can data assimilation do ?

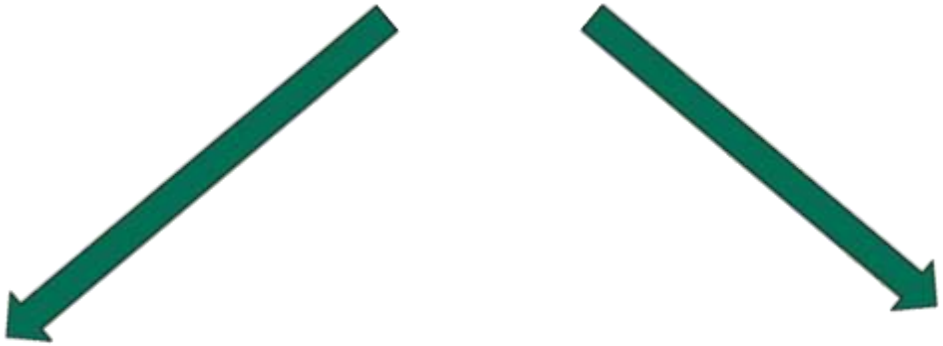


Flowchart of Data Assimilation



Data Assimilation arrests model divergence.

DATA ASSIMILATION



Finding maximum likelihood
(using Bayes' Theorem)

Minimize the cost function
(Least square approach)

WHAT IS BAYES' THEOREM ?

Bayes' Theorem

$$p(A \mid B) = \frac{p(B \mid A)p(A)}{p(B)}$$

P(A|B) = Probability of finding A given B
P(B|A) = Probability of finding B given A
P(A) = Probability of A with no knowledge of B
P(B) = Probability of B with no knowledge of A.

Did you ever bet on horses ?

Total Number of Races	12
Fleetfloot winning	7
Bolt winning	5

Probability of Bolt winning = 5/12 = 41.7%

Probability of Fleetfoot winning = 7/12 = 58.3%

Now let's add a new factor into the calculation. It turns out that on 3 of Bolt's previous 5 wins, it had rained heavily before the race. However, it had rained only once on any of the days that he lost. It appears, therefore, that Bolt is a horse who likes 'soft going', as the bookies say. On the day of the race in question, it is raining.

Given this new information (raining), what is the probability of Bolt winning ?

Ref : http://www.kevinboone.net/bayes.html

	It's raining	Not raining
Bolt winning	3	2
Bolt losing	1	6

What we need to know is the probability of Bolt winning, ***given that it is raining*** **?**

Like any other probability, we calculate it by dividing the number of times something happened, by the number of times if could have happened.
We know that Bolt won on 3 occasions on which it rained, and there were 4 rainy days in total.
So Bolt's probability of winning, given that it is now raining, is 3 / 4, or 0.75, or 75%.

The probability shifts from 41.7% to 75%.

This is important information if you plan to bet — **if it is raining you should back Bolt; if it is not, you should back Fleetfoot**.

Revisiting Bayes' Theorem

p(A|B) = p(B|A) p(A) / p(B)

P(A|B) = Probability of finding A given B
P(B|A) = Probability of finding B given A
P(A) = Probability of A with no knowledge of B
P(B) = Probability of B with no knowledge of A.

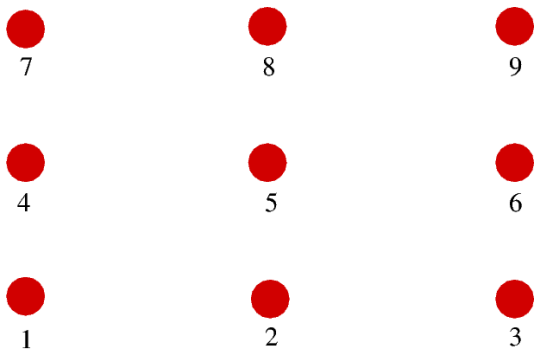
P(A|B) = Probability of Bolt winning when it rains
P(B|A) = Probability of raining when Bolt wins = 3/5
P(A) = Probability of Bolt winning = 5/12
P(B) = Probability of raining = 4/12

$$p(A|B) = \left(\frac{3}{5} \times \frac{5}{12}\right) \div \frac{4}{12} = \frac{3}{4}$$

BASICS

$$x^b_{k+1} = Mx^b_k + \varepsilon^b$$

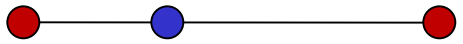
$$y_{k+1} = Hx^b_{k+1} + \varepsilon^o$$



What is x ?

What is y ?

What is H ?



What are the error characteristics ?

- Unbiased model and observation error i.e., $\langle \varepsilon^b \rangle = \langle \varepsilon^o \rangle = 0$
- Model and observation error are uncorrelated i.e., $\langle \varepsilon^b \varepsilon^{o^T} \rangle = 0$
- Non-trivial error covariances i.e., $\langle \varepsilon^b \varepsilon^{b^T} \rangle = B, \quad \langle \varepsilon^o \varepsilon^{o^T} \rangle = R$

What is B and R ?

$$p(x \mid y) \propto p(y \mid x)p(x)$$

Given observations, what is the best estimate of the state x.

$$p(x) \propto \exp\left(-\frac{1}{2}(x-x^b)B^{-1}(x-x^b)^T\right) \Leftarrow$$ B is Gaussian

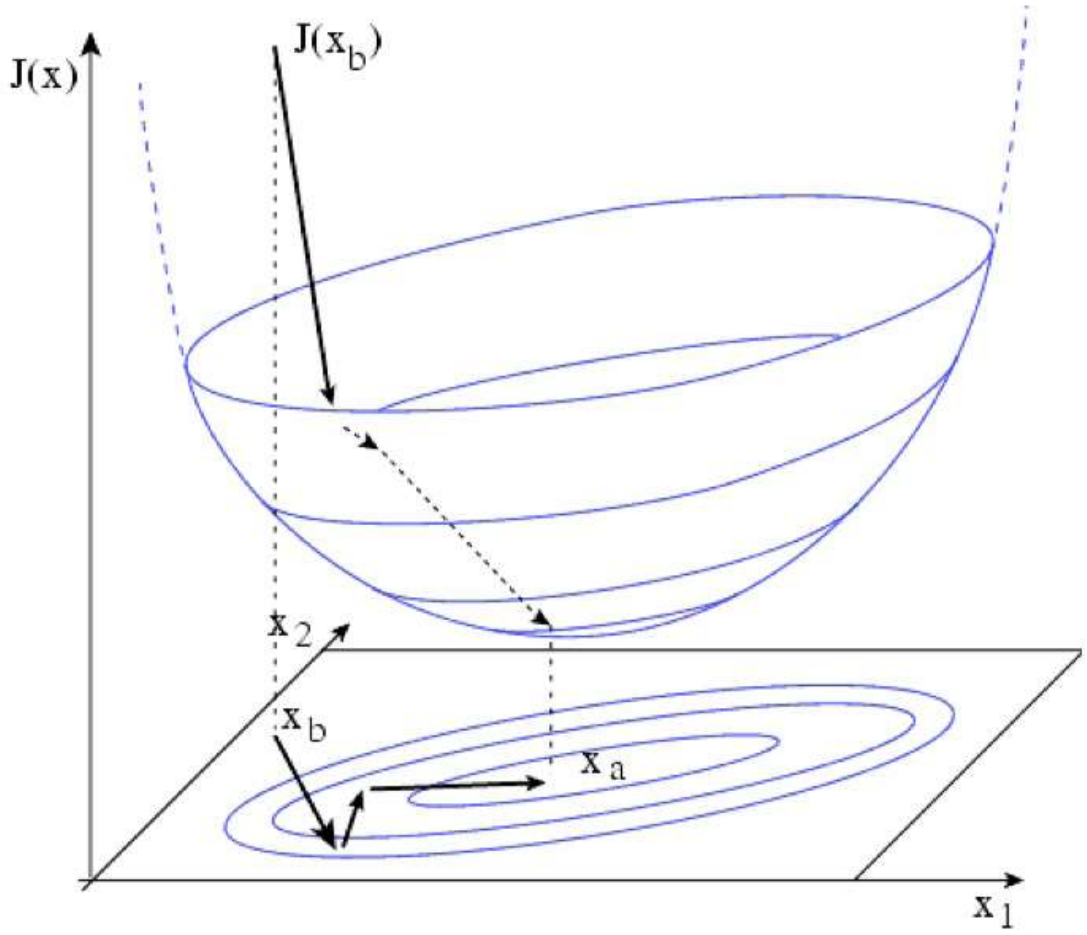
$$p(y \mid x) \propto \exp\left(-\frac{1}{2}(Hx-y)R^{-1}(Hx-y)^T\right) \Leftarrow$$ R is Gaussian

$$p(x \mid y) \propto \exp\left(-\frac{1}{2}J(x)\right) \quad \text{where}$$

$$J(x) = (x-x^b)B^{-1}(x-x^b)^T + (Hx-y)R^{-1}(Hx-y)^T$$

Maximizing $p(x \mid y)$ is same as **Minimizing** $J(x)$

$$x^a = x^b + BH^T\left(HBH^T + R\right)^{-1}\left(y - Hx^b\right)$$



The cost function is parabolic and the minimization is done using steepest descent.

POPULAR DATA ASSIMILATION METHODS

- KALMAN FILTER -- B evolves according to model dynamics.
- 3D VAR – B is stationary.
- 4D VAR – B evolves within the time window of cost function minimization.
- ENSEMBLE BASED KALMAN FILTER -- B is estimated from the ensembles.

What is the relative significance of B & R ?

$$x^a = x^b + BH^T\left(HBH^T + R\right)^{-1}\left(y - Hx^b\right)$$

Let's estimate the temperature of Hyderabad.

Given

$$x^b = 31.0, \sigma_b^2 = 2$$
$$y_0 = 30.0, \sigma_0^2 = 1$$

In this case, $H = 1, R = \sigma_o^2, B = \sigma_b^2$

$$x^a = x^b + \sigma_b^2\left(\sigma_b^2 + \sigma_o^2\right)^{-1}\left(y_0 - x^b\right)$$

$$\Rightarrow x^a = \frac{\sigma_0^2}{\sigma_0^2 + \sigma_b^2} x^b + \frac{\sigma_b^2}{\sigma_0^2 + \sigma_b^2} y_0$$

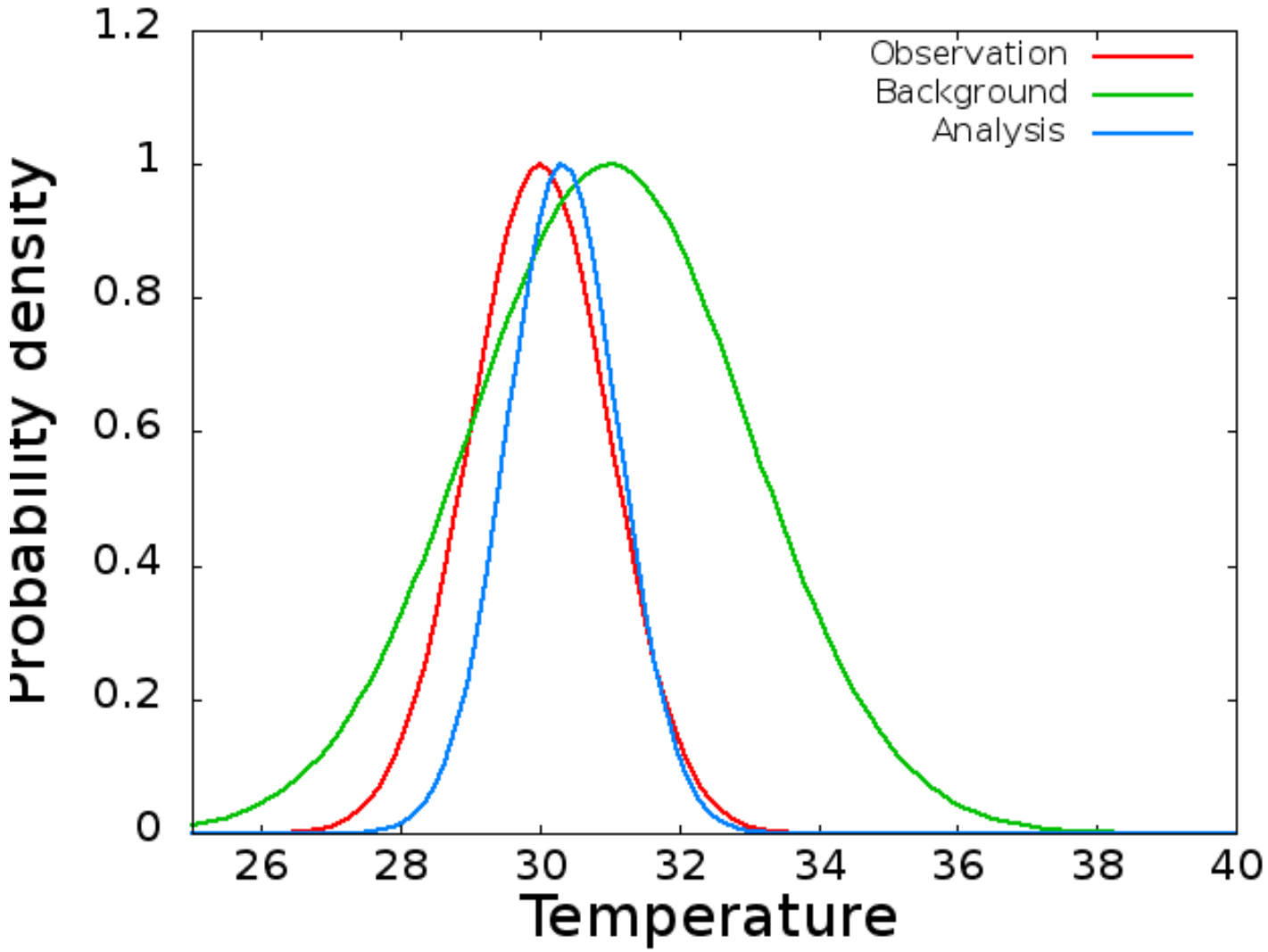
$$x^a = 30.33, \sigma_a^2 = 0.8$$

If $\sigma_b >> \sigma_0$

$x^a \approx y_0$

If $\sigma_0 >> \sigma_b$

$x^a \approx x_b$



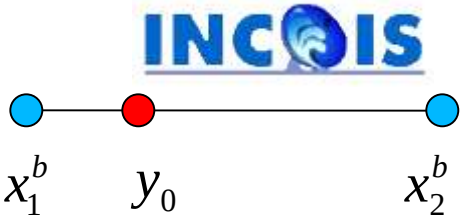
$x^b = 31.0, \sigma_b^2 = 2$

$y_0 = 30.0, \sigma_0^2 = 1$

$x^a = 30.33, \sigma_a^2 = 0.8$

Some more exercises

$$x^a = x^b + BH^T\left(HBH^T + R\right)^{-1}\left(y - Hx^b\right)$$



Suppose we observe a point in between two grid points.

$$Hx^b = \alpha x_1^b + (1-\alpha)x_2^b; \quad 0 \le \alpha \le 1$$

Assume

$$B = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} = \begin{bmatrix} \sigma_b^2 & \mu\sigma_b^2 \\ \mu\sigma_b^2 & \sigma_b^2 \end{bmatrix}; \qquad R = \sigma_0^2$$

$$\begin{pmatrix} x_1^a \\ x_2^a \end{pmatrix} = \begin{pmatrix} x_1^b \\ x_2^b \end{pmatrix} + \sigma_b^2 \begin{pmatrix} \alpha + \mu(1-\alpha) \\ \mu\alpha + (1-\alpha) \end{pmatrix} \frac{y_0 - \left[\alpha x_1^b + (1-\alpha)x_2^b\right]}{\left[\alpha^2 + 2\alpha(1-\alpha)\mu + (1-\alpha)^2\right]\sigma_b^2 + \sigma_0^2}$$

Case 1: No cross-correlation between two grid points, $\mu = 0$ and $\alpha = 1$

$$\begin{pmatrix} x_1^a \\ x_2^a \end{pmatrix} = \begin{pmatrix} x_1^b \\ x_2^b \end{pmatrix} + \sigma_b^2 \begin{pmatrix} \alpha + \mu(1-\alpha) \\ \mu\alpha + (1-\alpha) \end{pmatrix} \frac{y_0 - \left[\alpha x_1^b + (1-\alpha) x_2^b\right]}{\left[\alpha^2 + 2\alpha(1-\alpha)\mu + (1-\alpha)^2\right]\sigma_b^2 + \sigma_0^2}$$

$$\Rightarrow \begin{pmatrix} x_1^a \\ x_2^a \end{pmatrix} = \begin{pmatrix} x_1^b \\ x_2^b \end{pmatrix} + \sigma_b^2 \begin{pmatrix} 1 \\ 0 \end{pmatrix} \frac{y_0 - x_1^b}{\sigma_b^2 + \sigma_0^2}$$

$$x_1^a = \frac{\sigma_0^2}{\sigma_0^2 + \sigma_b^2} x_1^b + \frac{\sigma_b^2}{\sigma_0^2 + \sigma_b^2} y_0$$

$$x_2^a = x_2^b$$

The analysis at grid point 2 is equal to the background. Observation had no effect.

Case 2: $\alpha = 1, \quad \mu \neq 0$

$$\begin{pmatrix} x_1^a \\ x_2^a \end{pmatrix} = \begin{pmatrix} x_1^b \\ x_2^b \end{pmatrix} + \sigma_b^2 \begin{pmatrix} \alpha + \mu(1-\alpha) \\ \mu\alpha + (1-\alpha) \end{pmatrix} \frac{y_0 - \left[\alpha x_1^b + (1-\alpha) x_2^b\right]}{\left[\alpha^2 + 2\alpha(1-\alpha)\mu + (1-\alpha)^2\right]\sigma_b^2 + \sigma_0^2}$$

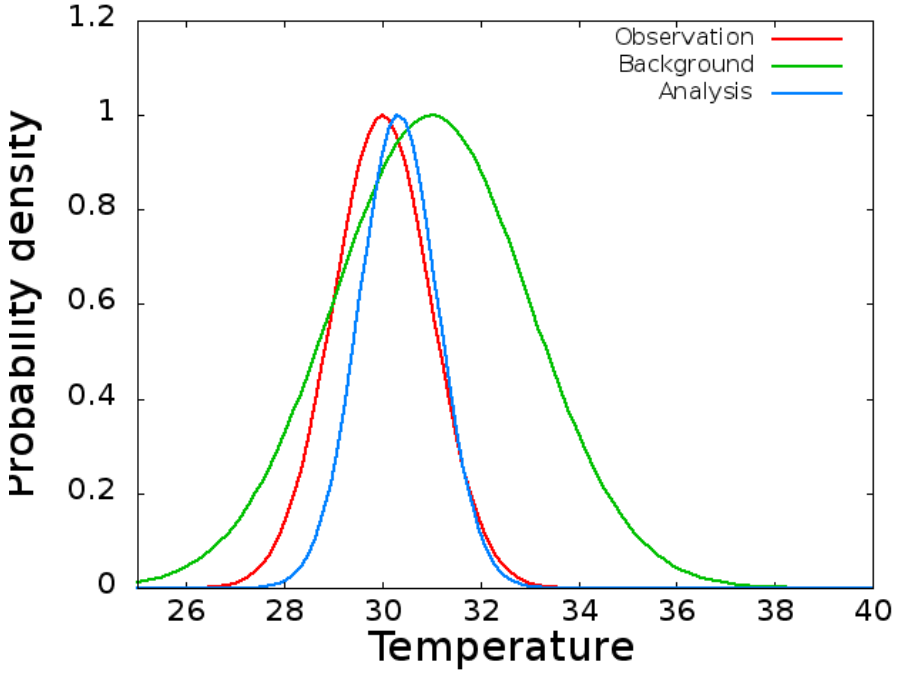
$$\Rightarrow \begin{pmatrix} x_1^a \\ x_2^a \end{pmatrix} = \begin{pmatrix} x_1^b \\ x_2^b \end{pmatrix} + \sigma_b^2 \begin{pmatrix} 1 \\ \mu \end{pmatrix} \frac{y_0 - x_1^b}{\sigma_b^2 + \sigma_0^2}$$

$$x_1^a = \frac{\sigma_0^2}{\sigma_0^2 + \sigma_b^2} x_1^b + \frac{\sigma_b^2}{\sigma_0^2 + \sigma_b^2} y_0$$

$$x_2^a = x_2^b + \mu\sigma_b^2 \frac{y_0 - x_1^b}{\sigma_0^2 + \sigma_b^2}$$

Now the solution at grid point 2 is influenced by the observation. The role of Background error covariance is to spread information from one grid point to the other.

PRACTICAL ISSUES



$$x^a = \frac{\sigma_0^2}{\sigma_0^2 + \sigma_b^2} x^b + \frac{\sigma_b^2}{\sigma_0^2 + \sigma_b^2} y_0$$

If $\sigma_b >> \sigma_0$

$x^a \approx y_0$

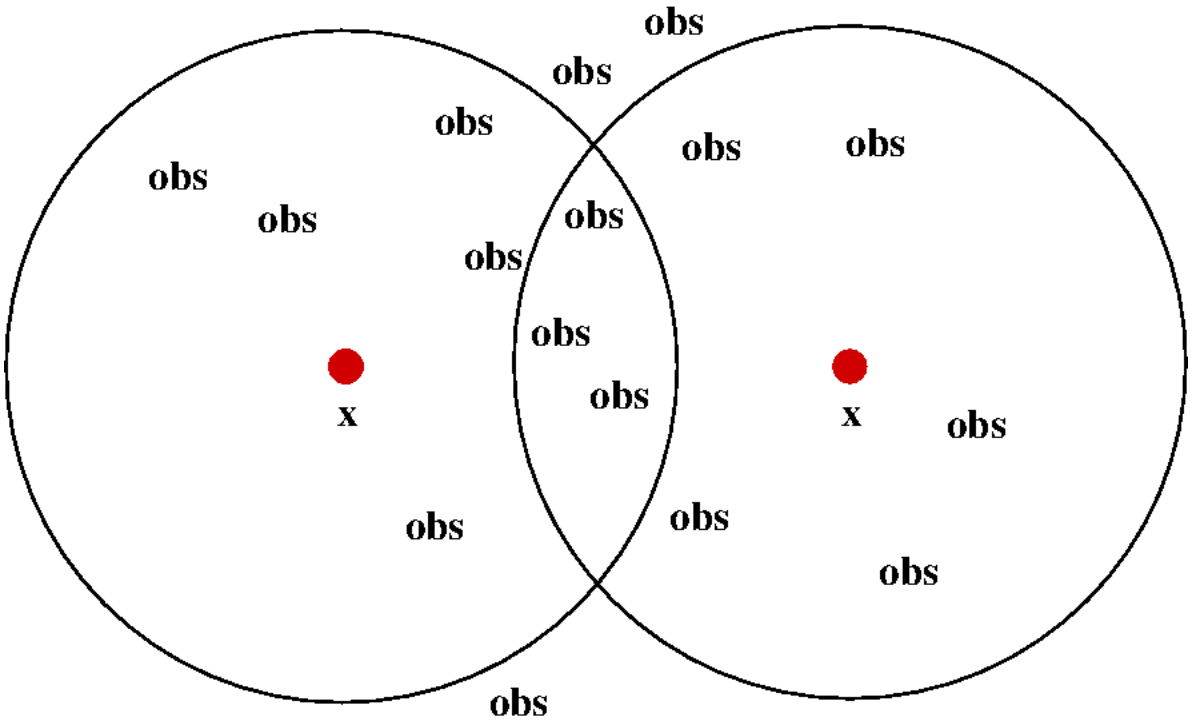
If $\sigma_0 >> \sigma_b$

$x^a \approx x_b$

COVARIANCE INFLATION IS NECESSARY !!!

Assimilating distant observations leads to spurious correlations

Idea of Localization



PRACTICAL APPLICATIONS IN INCOIS

OBSERVATIONS

Assimilated Variables

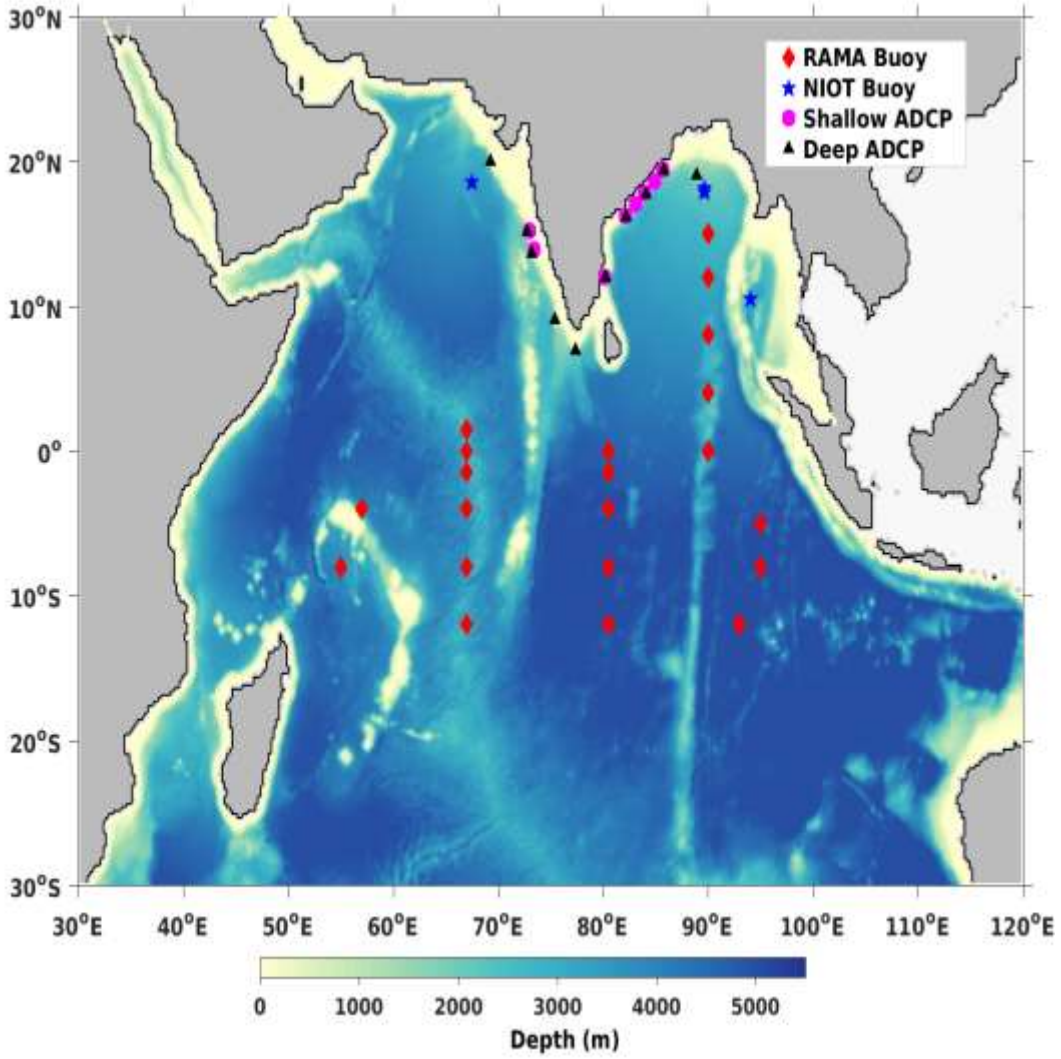
1. In-situ Temperature
2. Salinity Profiles (RAMA moorings, NIOT buoys and Argo floats)
3. Sea surface temperature (Satellite track data : AMRSE)

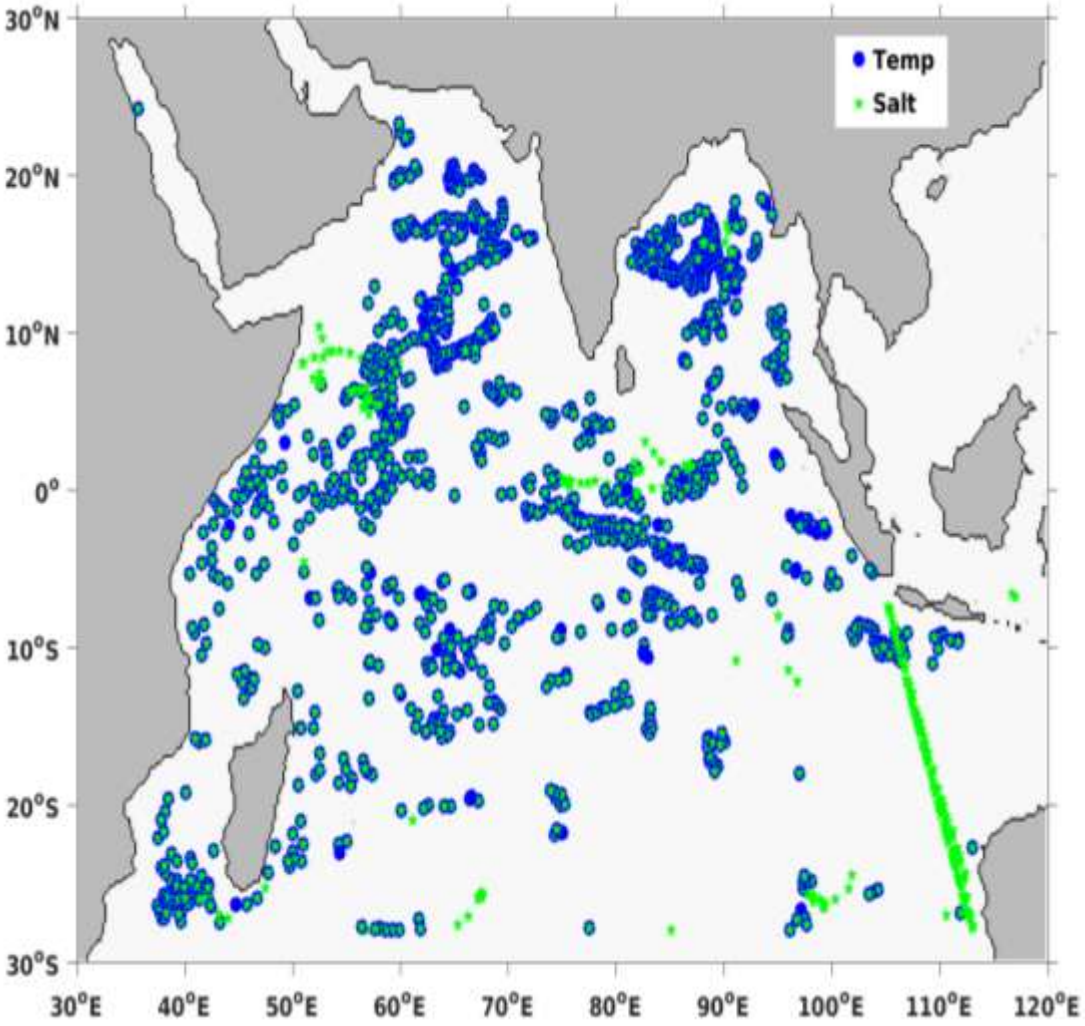
Independent Variables

1. Sea level anomaly
2. Sea Surface salinity
3. U,V Currents

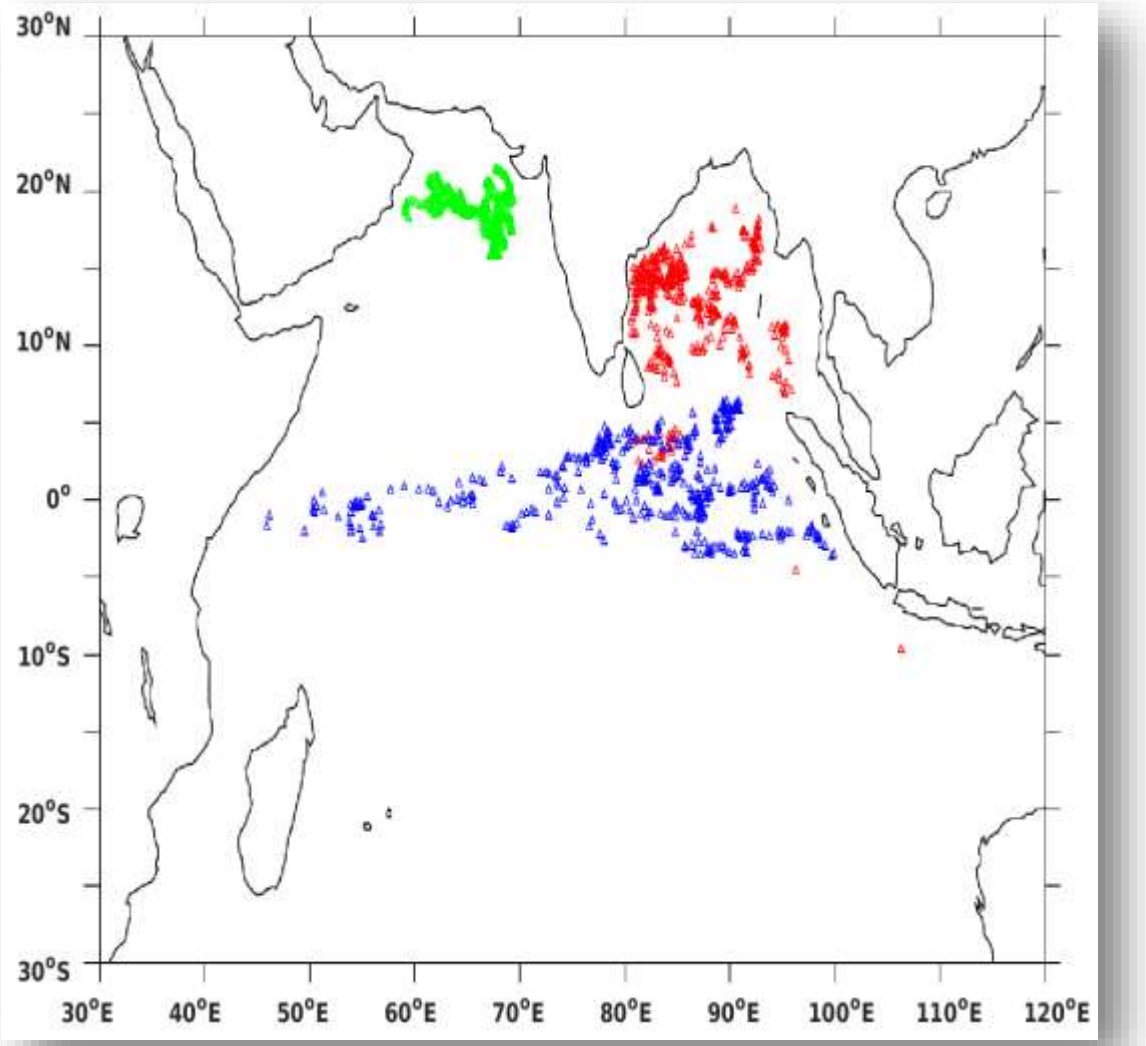
Validations and Comparisons were made with respect to both assimilated (dependent) variables and Independent variables

Variable	Assimilated	Validation / Comparison
SST	AMRSE satellite track data	AVHRR
SLA	-	AVISO
U V Currents	-	OSCAR, ADCP
Temperature	In-situ profiles	RAMA mooring, NIOT Buoys
Salinity	In-situ profiles	RAMA mooring, NIOT Buoys


RAMA Buoy
NIOT Buoy
Shallow ADCP
Deep ADCP
30°N
20°N
10°N
0°
10°S
20°S
30°S
30°E
40°E
50°E
60°E
70°E
80°E
90°E
100°E
110°E
120°E
0
1000
2000
3000
4000
5000
Depth (m)

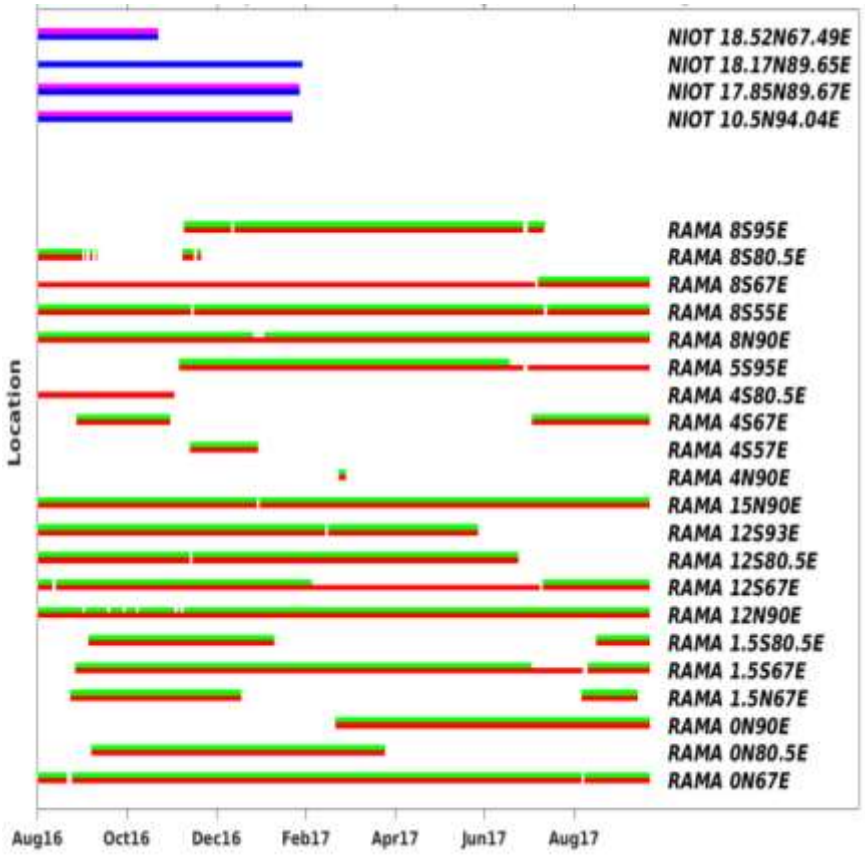

Temp
Salt
30°N
20°N
10°N
0°
10°S
20°S
30°S
30°E
40°E
50°E
60°E
70°E
80°E
90°E
100°E
110°E
120°E

ARGO FLOATS

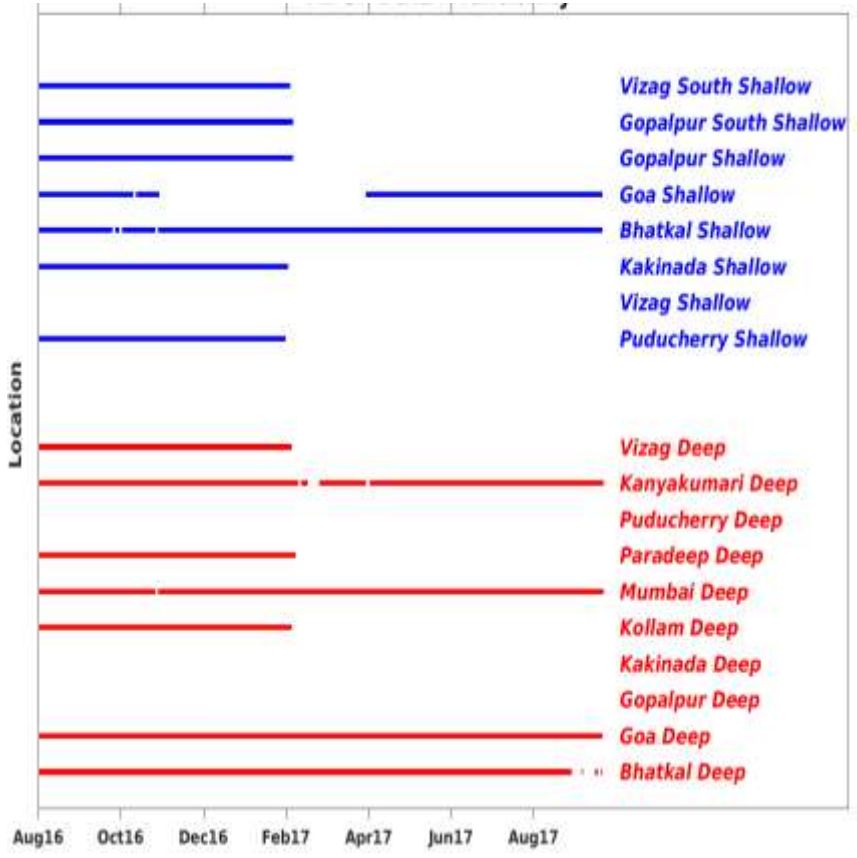


Daily pop-up of Argo floats in the Northern Indian Ocean

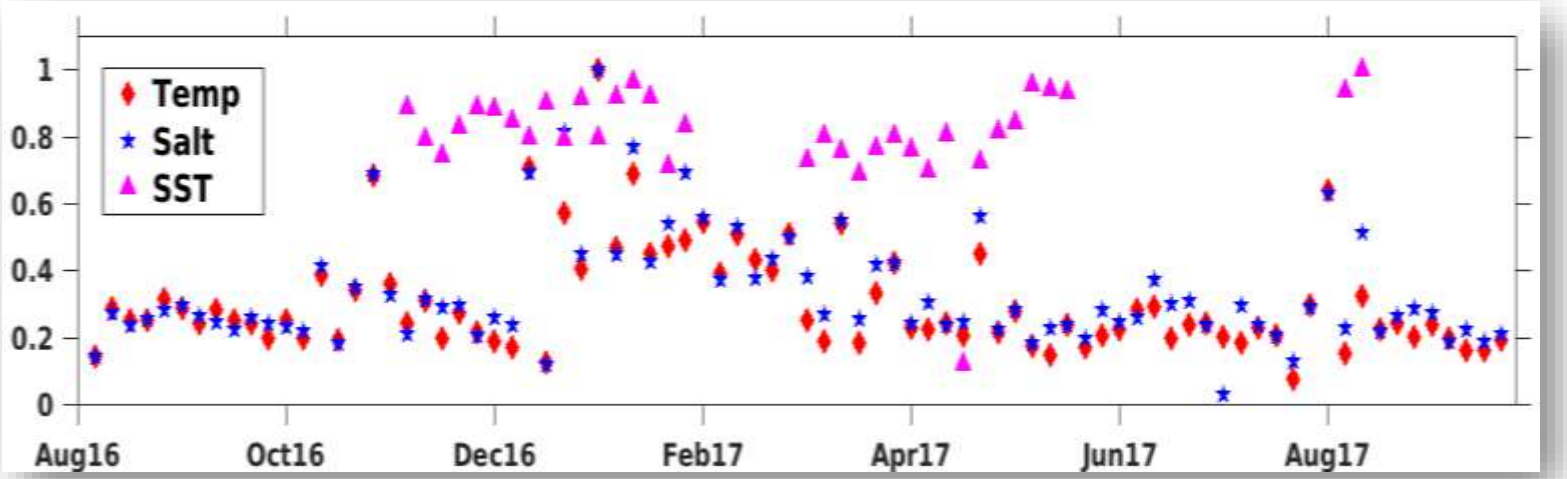
Temperature and Salinity Data availability



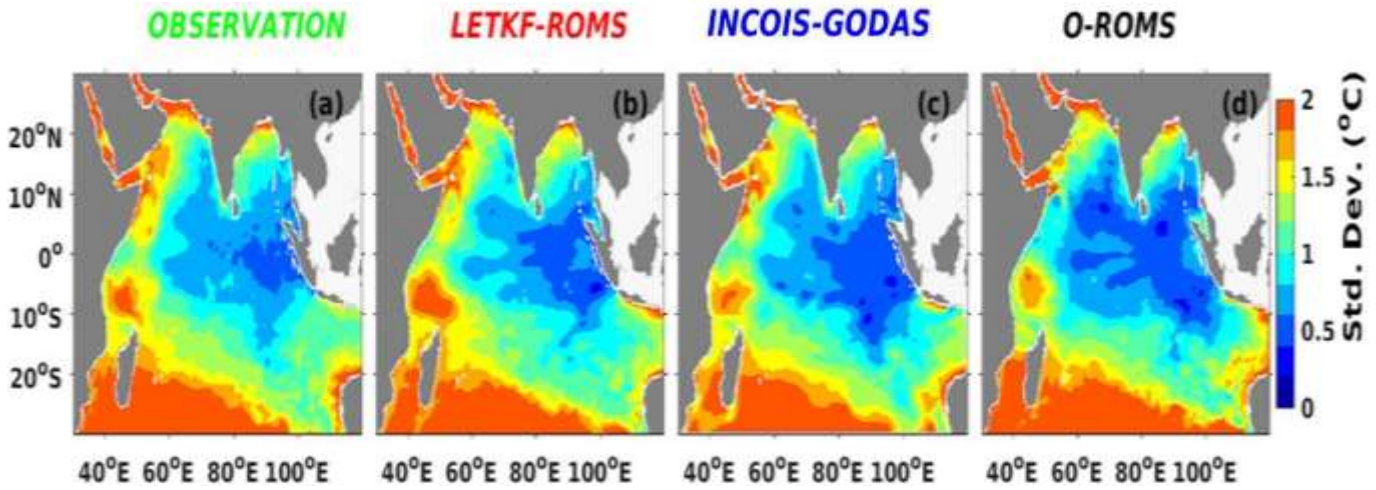
ADCP Data availability



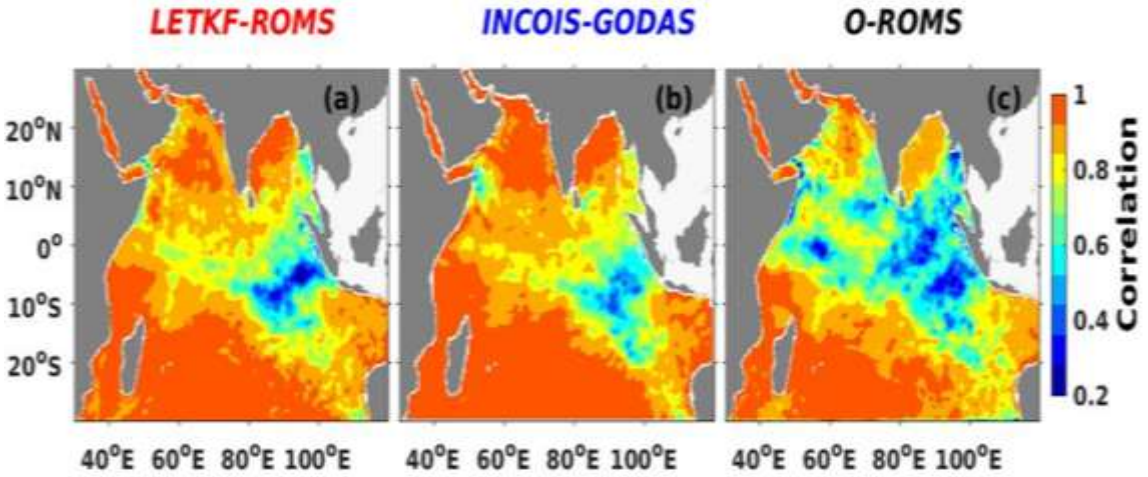
Assimilated Observations



SST

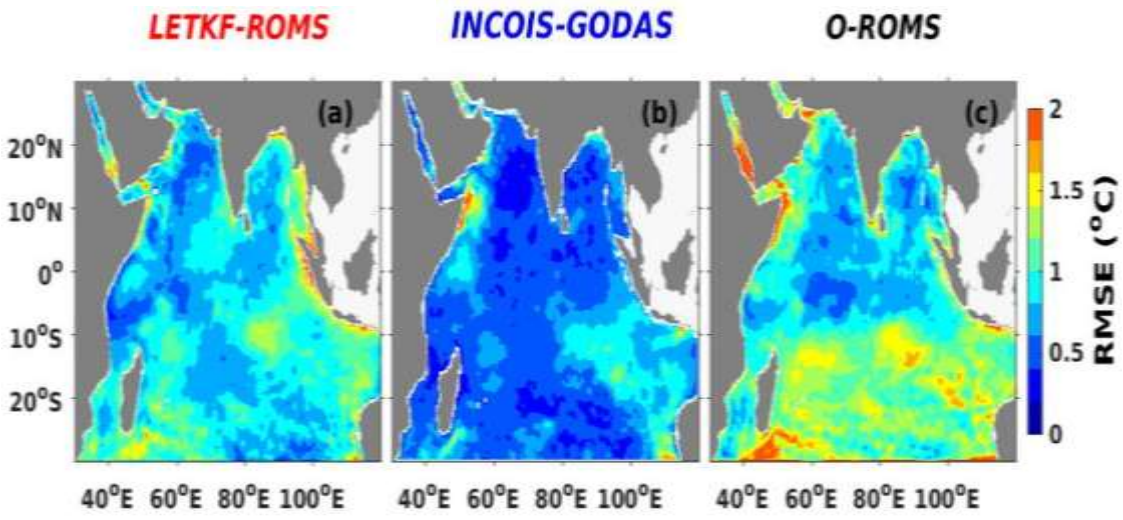


Standard Deviation of SST

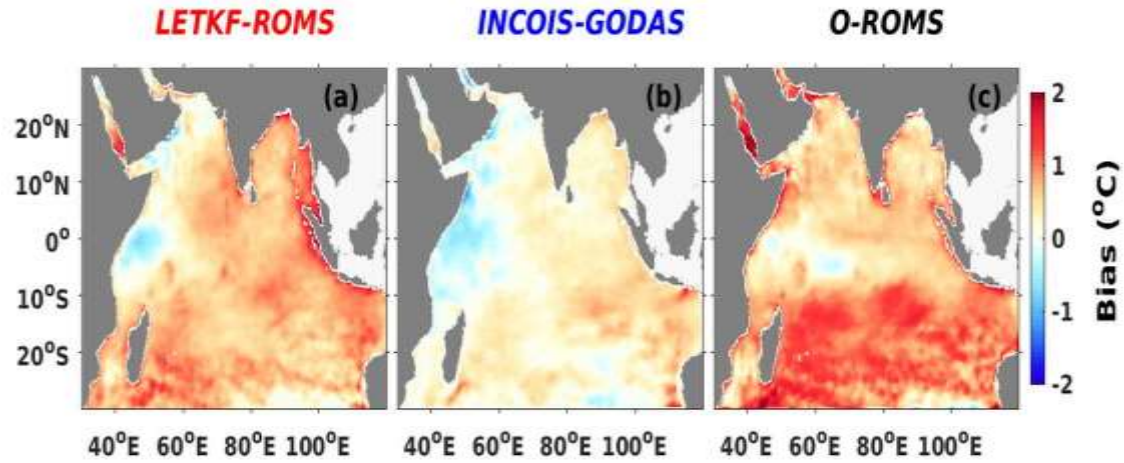


Correlation of SST with respect to AVHRR Data

SST

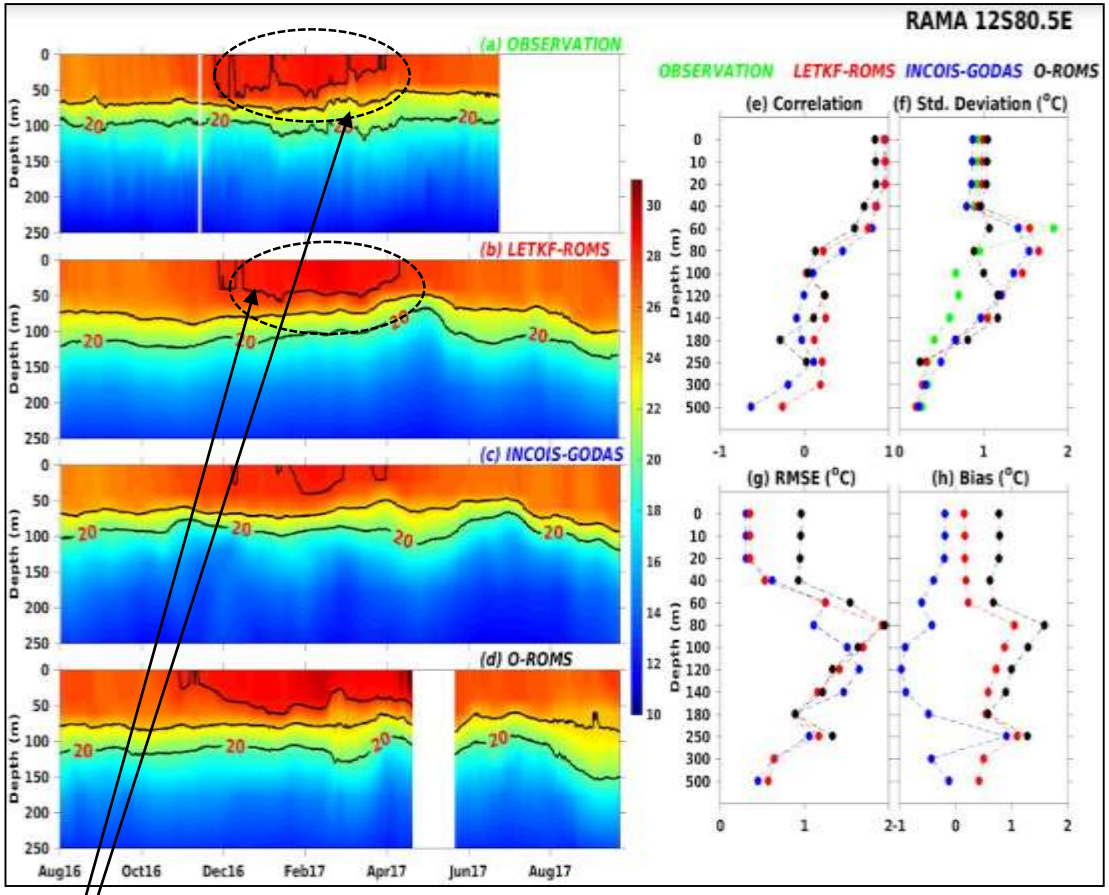


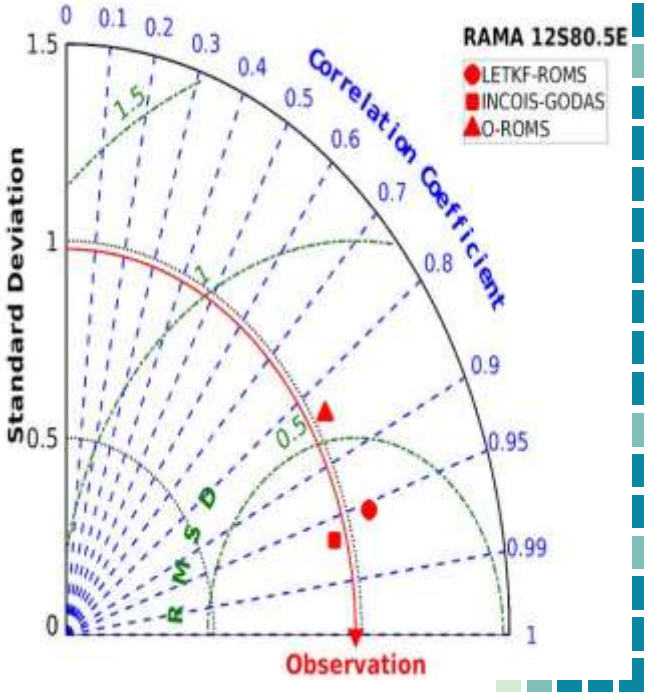
RMSE of SST with respect to AVHRR Data

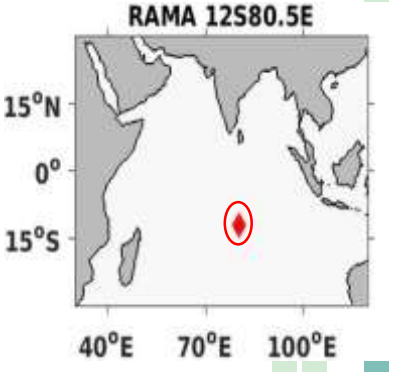


BIAS for SST with respect to AVHRR Data

Temperature







Temporal evolution of water in subsurface layers simulated by LETKF-ROMS is in good agreement with observation in this location as well.

TAKE HOME MESSAGE

- The truth is not known.
- Neither observation nor model is devoid of errors.
- Assimilate these two to get the best estimate.
- Estimating maximum likelihood = Minimizing cost function.
- The model error covariance propagates information from one place to another.
- Covariance inflation is necessary for Ensemble based schemes.
- Localize observations to get rid of spurious correlations.


THANK
YOU !