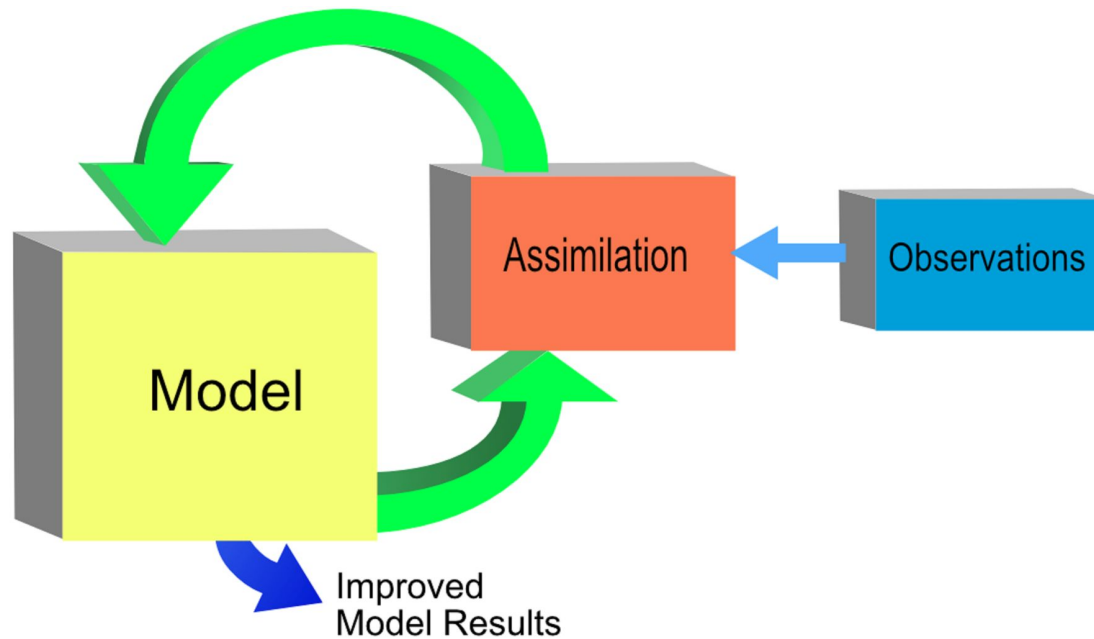


BASIC CONCEPTS OF DATA ASSIMILATION



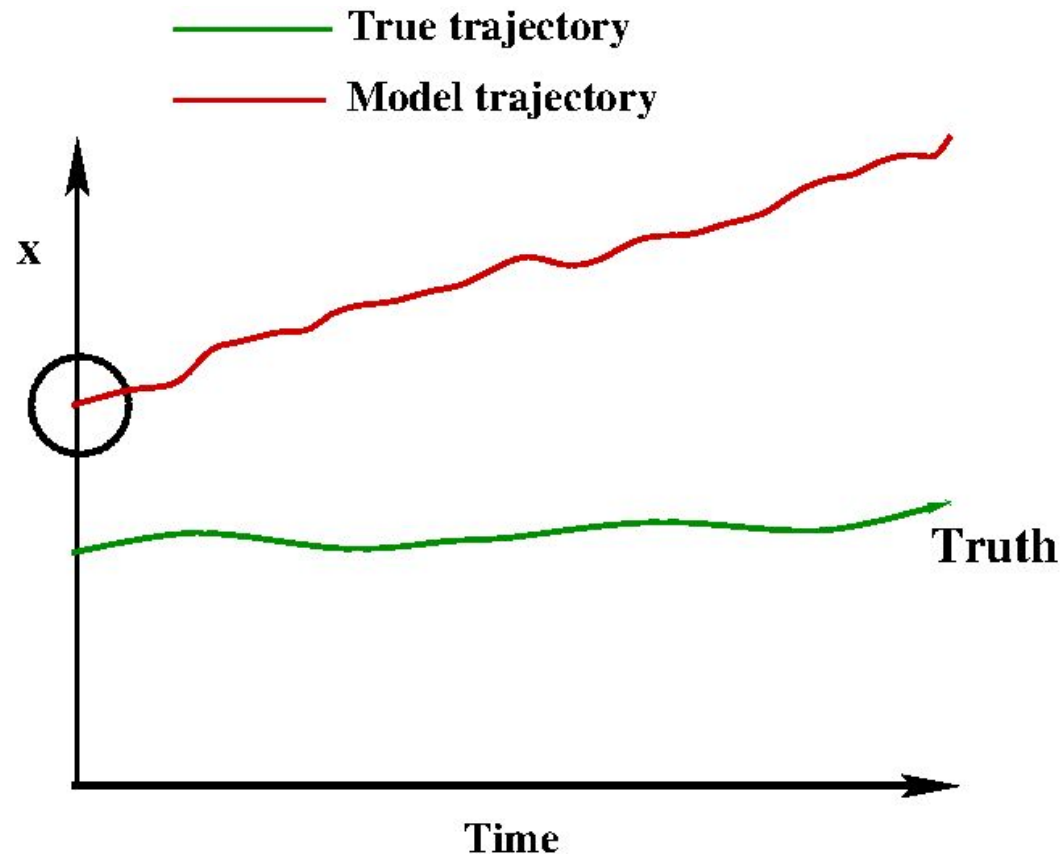
Training Course on Modelling for Ocean Forecasting and Process Studies during 6 – 10 December 2021

**ARYA
PAUL
INCOIS**

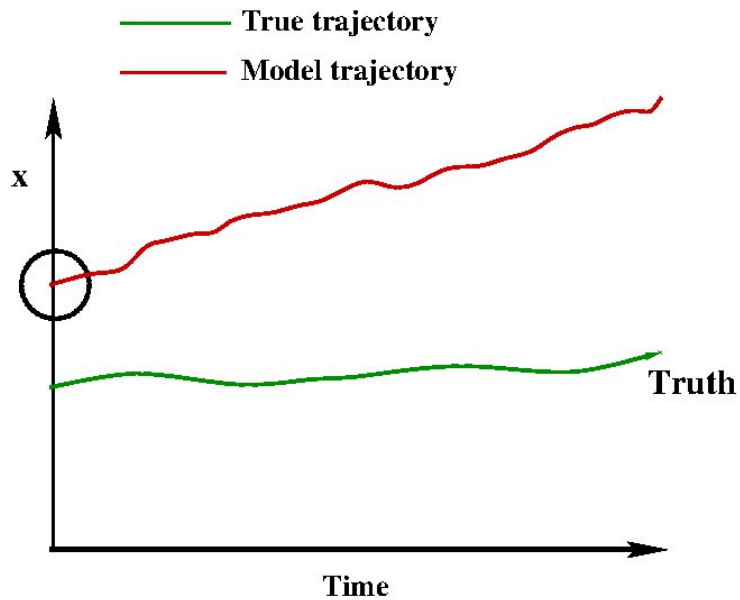
OUTLINE

- **Why** do we need data assimilation ?
- **What** is data assimilation ?
- **How** does it improve the estimate ?
- Practical Applications (if time permits).

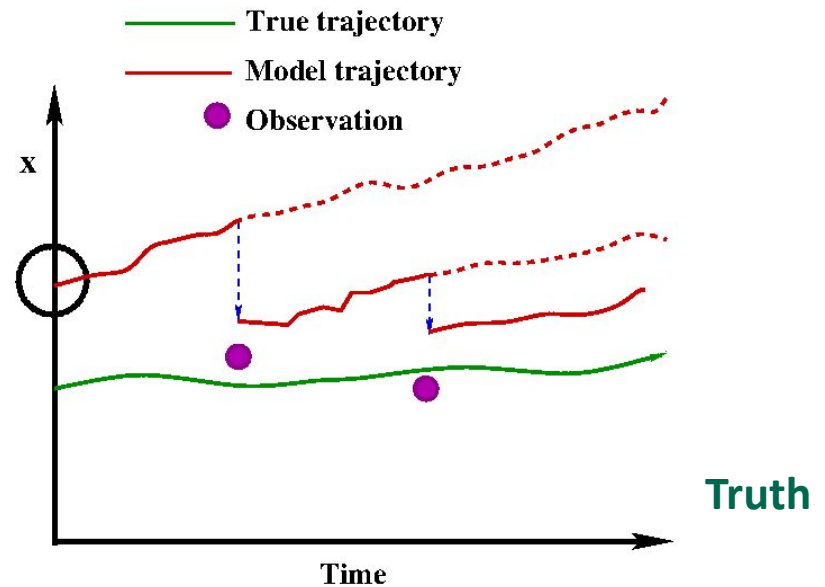
Why do we need Data Assimilation ?



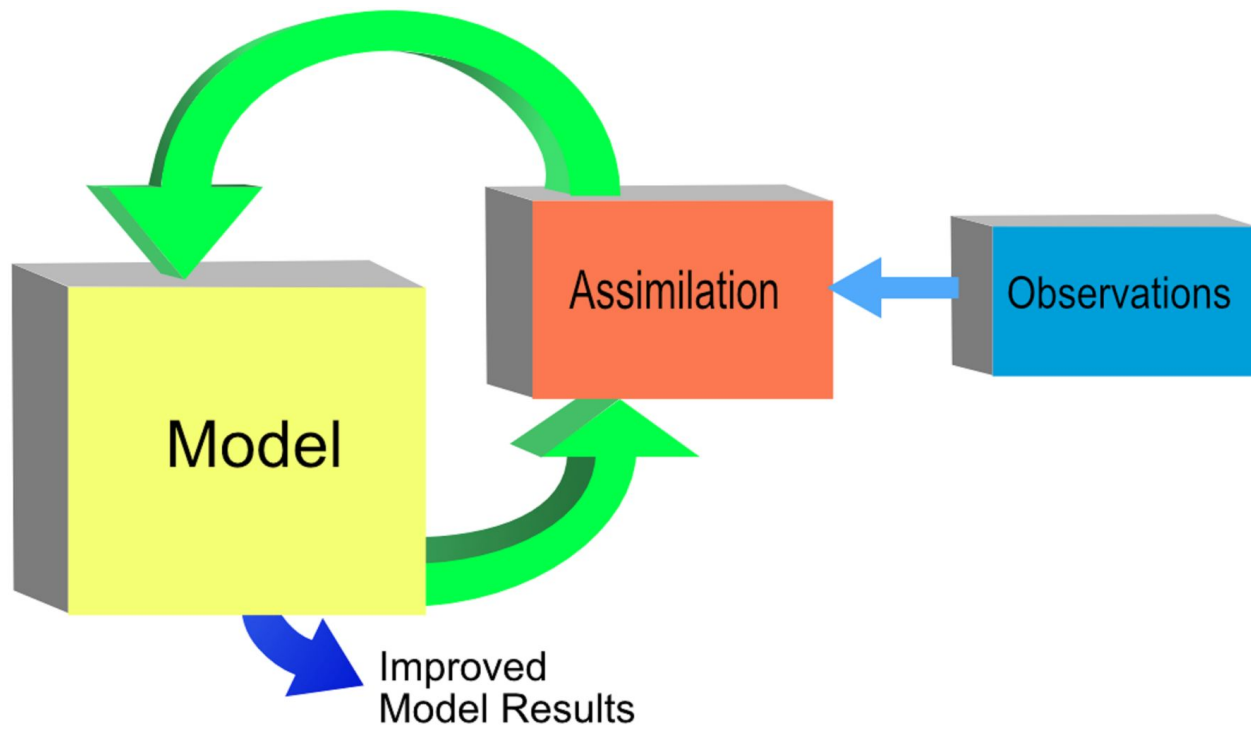
Flowchart of Data Assimilation



NO CORRECTION

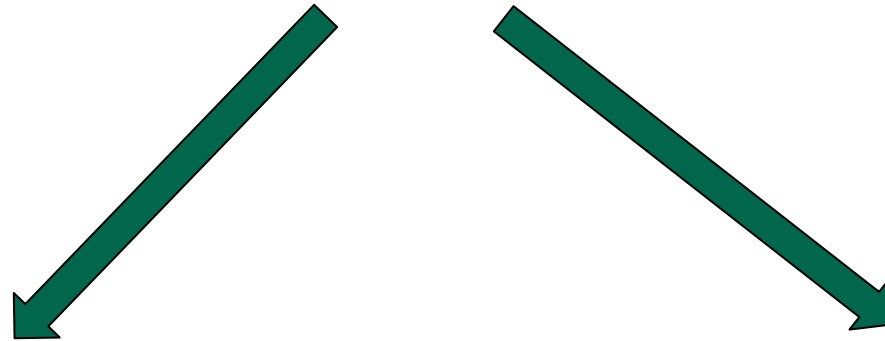


CORRECTION



What is Data Assimilation ?

DATA ASSIMILATION



Finding maximum likelihood
(using Bayes' Theorem)

Minimize the cost function
(Least square approach)

WHAT IS BAYES' THEOREM ?

What is Probability ?

How likely an event is likely to occur !!!

Bayes' Theorem

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

$P(A | B)$ = 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.

Did you ever bet on horses ?

Total Number of Races	12
Fleetfoot 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 it 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

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

Data Assimilation

The diagram illustrates the Data Assimilation equation, showing the relationship between the Analysis state vector, the Background state vector, the Model Error Covariance, the Projection Operator, the Observation Error Covariance, and the Observation.

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

Labels and their corresponding terms in the equation:

- Analysis state vector: x^a
- Background state vector: x^b
- Model Error Covariance: BH^T
- Projection Operator: H
- Observation Error Covariance: R
- Observation: y

What is a state vector ?

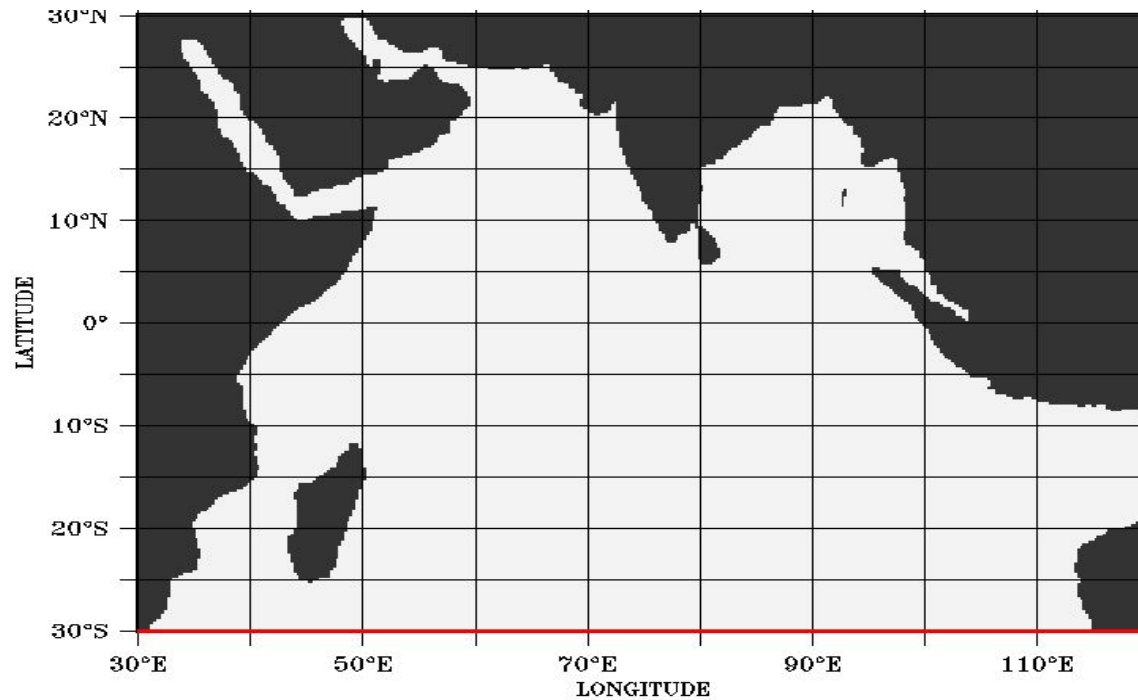
A **state variable** is one of the set of **variables** that are used to describe the mathematical "state" of a **dynamical system**.

A state vector is composed of all such state variables.

Example :-

In physical oceanography, a state vector is composed of temperature, salinity, horizontal velocities and sea surface height.

When you know the state vector at every location, you know the state of the ocean.



$$x = (T, S, u, v, SSH)$$

In numerical ocean modeling, a $\frac{1}{4}$ degree global model has a state vector whose length is

$$L = (90 \times 4) \times (360 \times 4) \times (\text{No. of vertical Layers}) \times 5 = 10^8 \longrightarrow \text{Huge Beast}$$

POPULAR DATA ASSIMILATION METHODS

- **KALMAN FILTER**
- **3D VAR**
- **4D VAR**
- **ENSEMBLE BASED METHODS**
- **NONLINEAR FILTERS**

What is the relative significance of B & R ?

$$\mathbf{x}^a = \mathbf{x}^b + \mathbf{B}\mathbf{H}^\top (\mathbf{H}\mathbf{B}\mathbf{H}^\top + \mathbf{R})^{-1} (\mathbf{y} - \mathbf{H}\mathbf{x}^b)$$

Let's estimate the temperature of Hyderabad.

Given

$$\begin{aligned} x^b &= 31.0, \sigma_b^2 = 2 \\ y_0 &= 30.0, \sigma_0^2 = 1 \end{aligned}$$

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

$$\begin{aligned} x^a &= x^b + \sigma_b^2 (\sigma_b^2 + \sigma_o^2)^{-1} (y_0 - x^b) \\ \Rightarrow x^a &= \frac{\sigma_0^2}{\sigma_0^2 + \sigma_b^2} x^b + \frac{\sigma_b^2}{\sigma_0^2 + \sigma_b^2} y_0 \end{aligned}$$

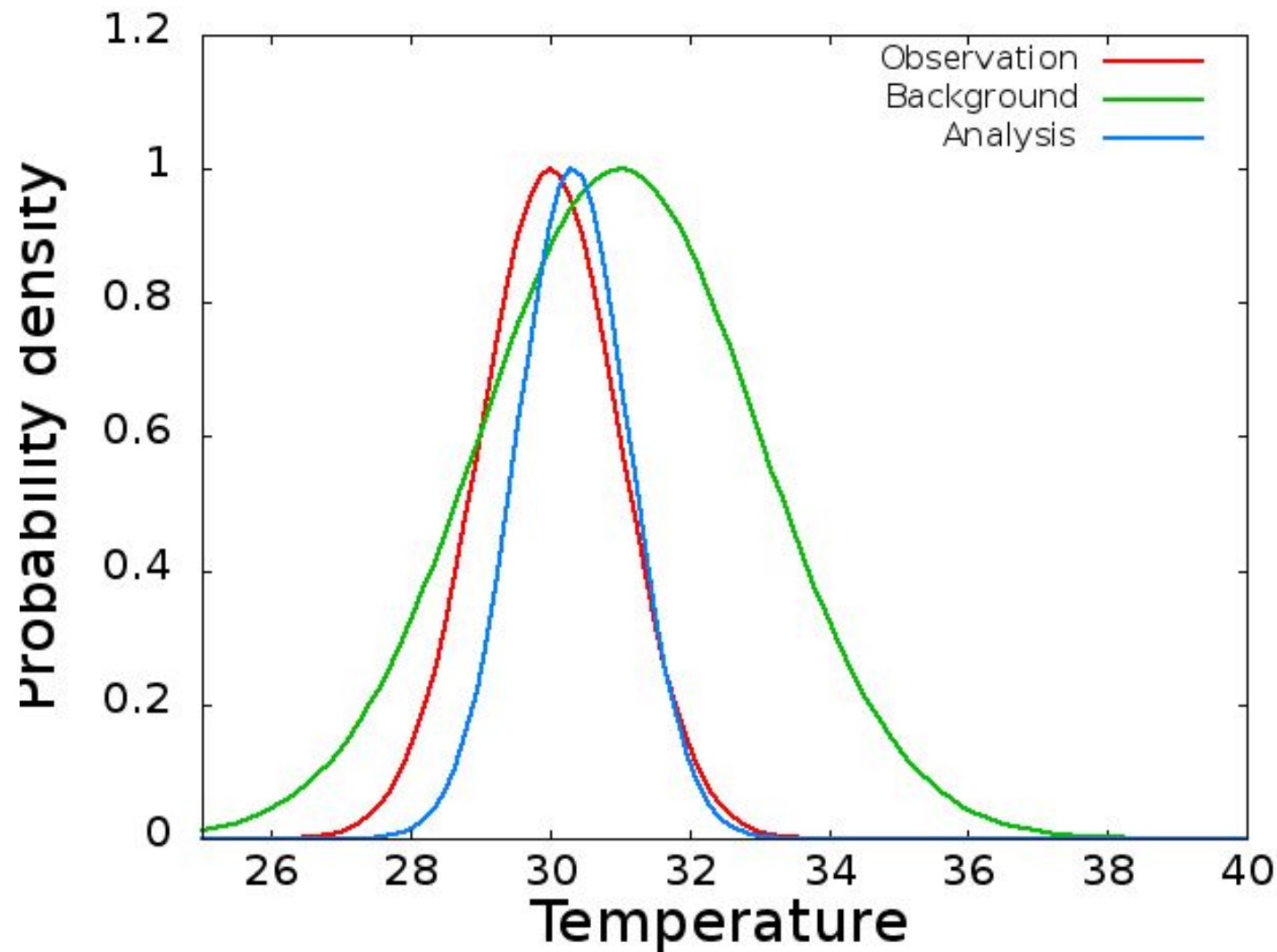
$$x^a = 30.33, \sigma_a^2 = 0.8$$

If $\sigma_b \gg \sigma_0$

$$x^a \approx y_0$$

If $\sigma_0 \gg \sigma_b$

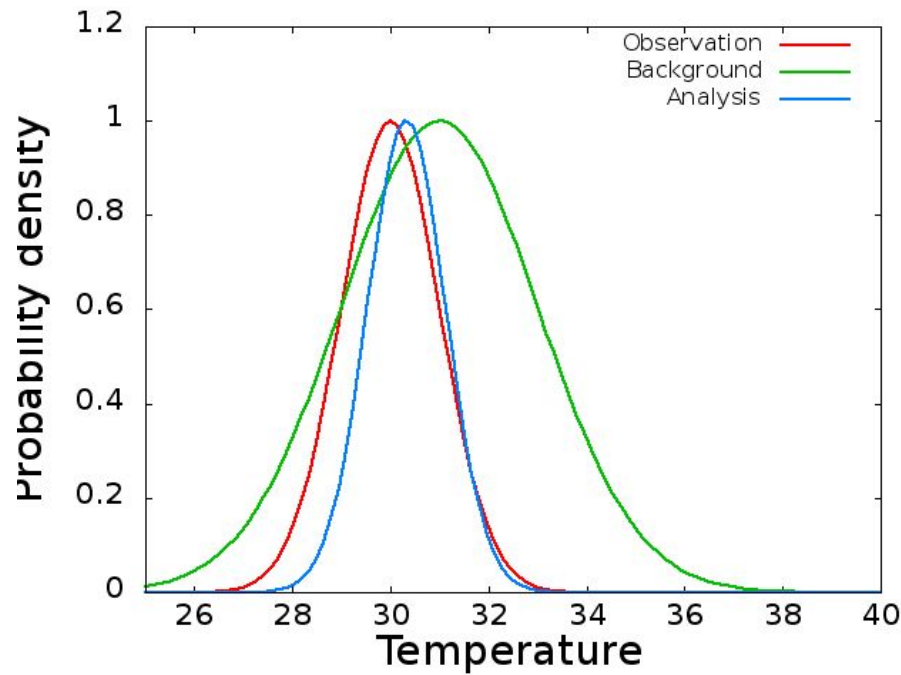
$$x^a \approx x_b$$



$$x^b = 31.0, \sigma_b^2 = 2$$
$$y_0 = 30.0, \sigma_0^2 = 1$$
$$x^a = 30.33, \sigma_a^2 = 0.8$$

IS IT ALWAYS GOOD TO GET AN IMPROVED CERTAINTY?

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 $\sigma_b \gg \sigma_0$

$$x^a \approx y_0$$

If $\sigma_0 \gg \sigma_b$

$$x^a \approx x_b$$

COVARIANCE INFLATION IS NECESSARY !!!

What is the role of B ??

B propagates information from one site to another !!!

$$\mathbf{x}^a = \mathbf{x}^b + \mathbf{B}\mathbf{H}^\top (\mathbf{H}\mathbf{B}\mathbf{H}^\top + \mathbf{R})^{-1} (\mathbf{y} - \mathbf{H}\mathbf{x}^b)$$



Suppose we observe a point in between two grid points.

$$\mathbf{H}\mathbf{x}^b = \alpha x_1^b + (1 - \alpha)x_2^b; \quad 0 \leq \alpha \leq 1$$

Assume

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

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

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

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

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

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

$$x_2^a = x_2^b$$

**The analysis at grid point 1 is same as the analysis of the previous example
The analysis at grid point 2 is equal to the background. Observation had no effect.**

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

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

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

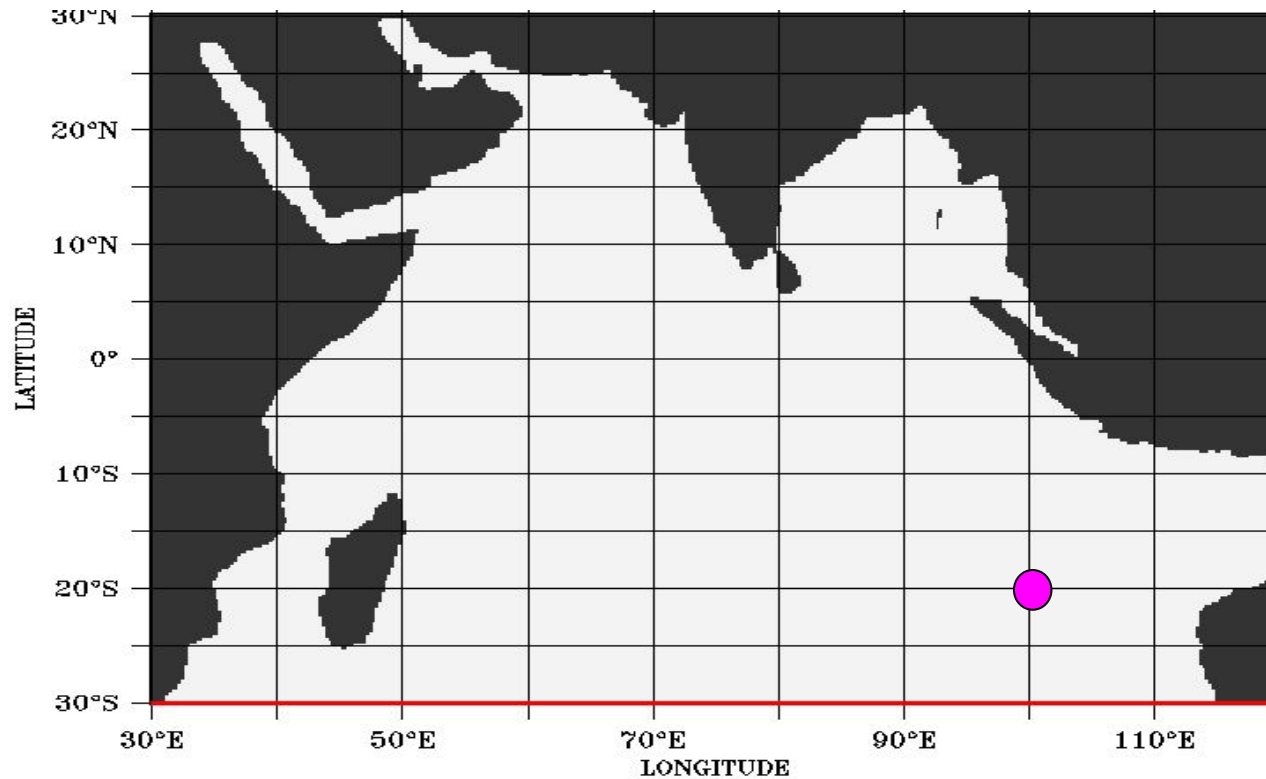
$$x_1^a = \frac{\sigma_0^2}{\sigma_0^2 + \sigma_b^2} x_1^b + \frac{\sigma_b^2}{\sigma_0^2 + \sigma_b^2} y_0$$

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

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.

ARE THERE DRAWBACKS OF PROPAGATING INFORMATION ?

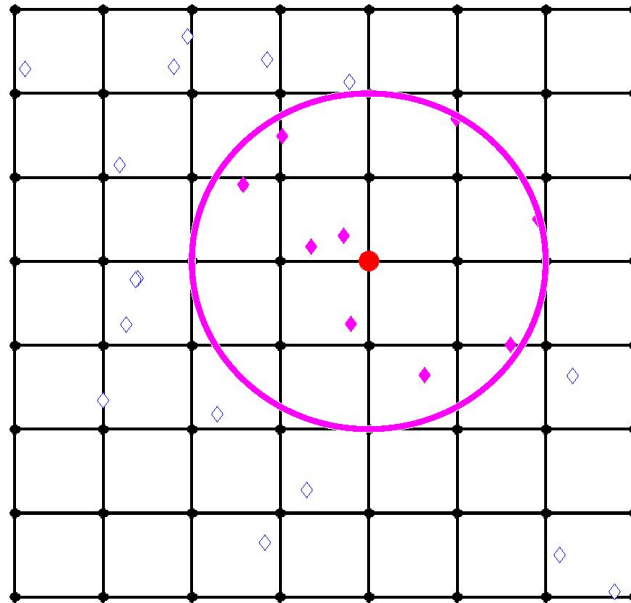
Idea of Localization



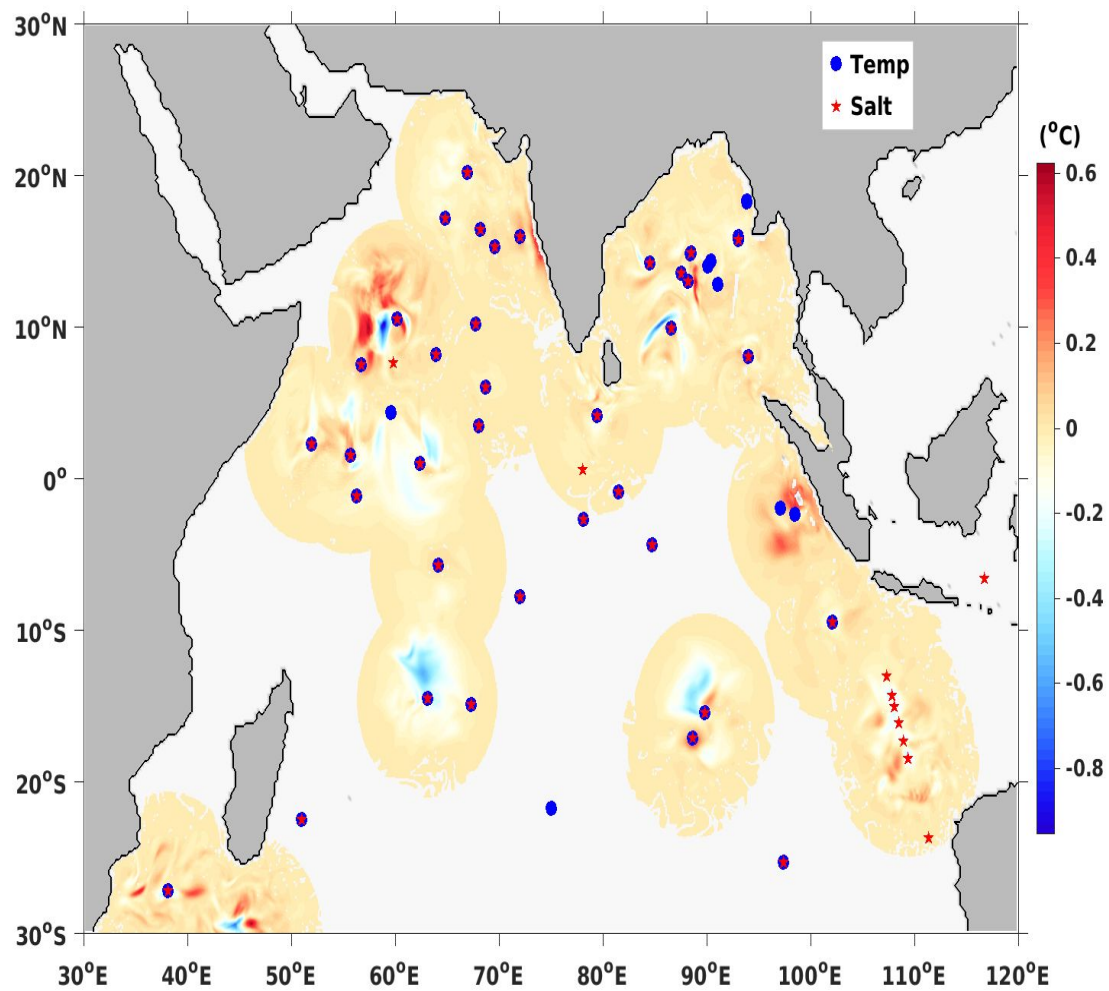
SHOULD THIS OBSERVATION INFLUENCE ALL GRID POINTS ?

Localization

Observation based **Localization**



- Data Assimilation is done in a local volume, choosing observations
- The state estimate is updated at the central grid red dot
- All observations (purple diamonds) within the local region are assimilated



TAKE HOME MESSAGE

- The truth is not known.
- Neither observation nor model is devoid of errors.
- Assimilate these two to get a best estimate.
- 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.

PRACTICAL APPLICATIONS IN INCOIS

OBSERVATIONS



```
graph TD; OBSERVATIONS --> Assimilated_Variables[Assimilated Variables]; OBSERVATIONS --> Independent_Variables[Independent Variables]; Assimilated_Variables --> A1[1. In-situ Temperature]; Assimilated_Variables --> A2[2. Salinity Profiles (RAMA moorings, NIOT buoys and Argo floats)]; Assimilated_Variables --> A3[3. Sea surface temperature (Satellite track data : AMRSE)]; Independent_Variables --> I1[1. Sea level anomaly]; Independent_Variables --> I2[2. Sea Surface salinity]; Independent_Variables --> I3[3. U,V Currents];
```

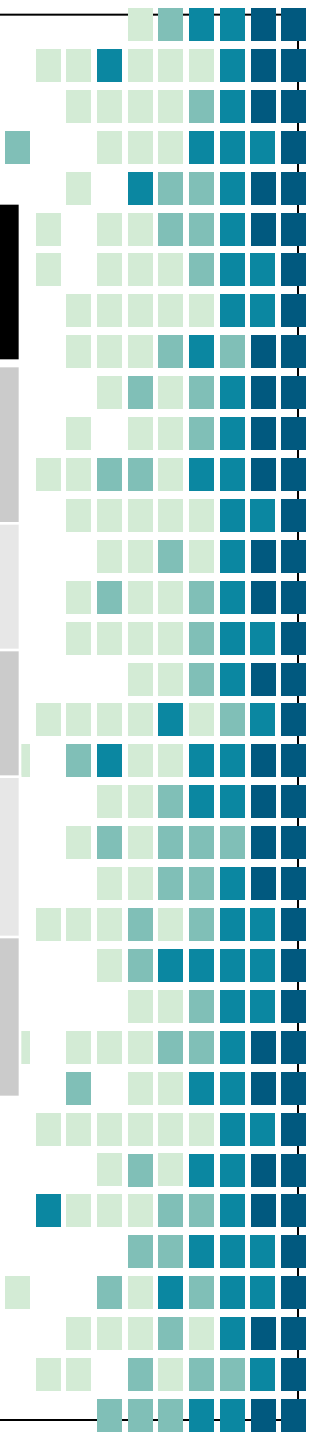
Assimilated Variables

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

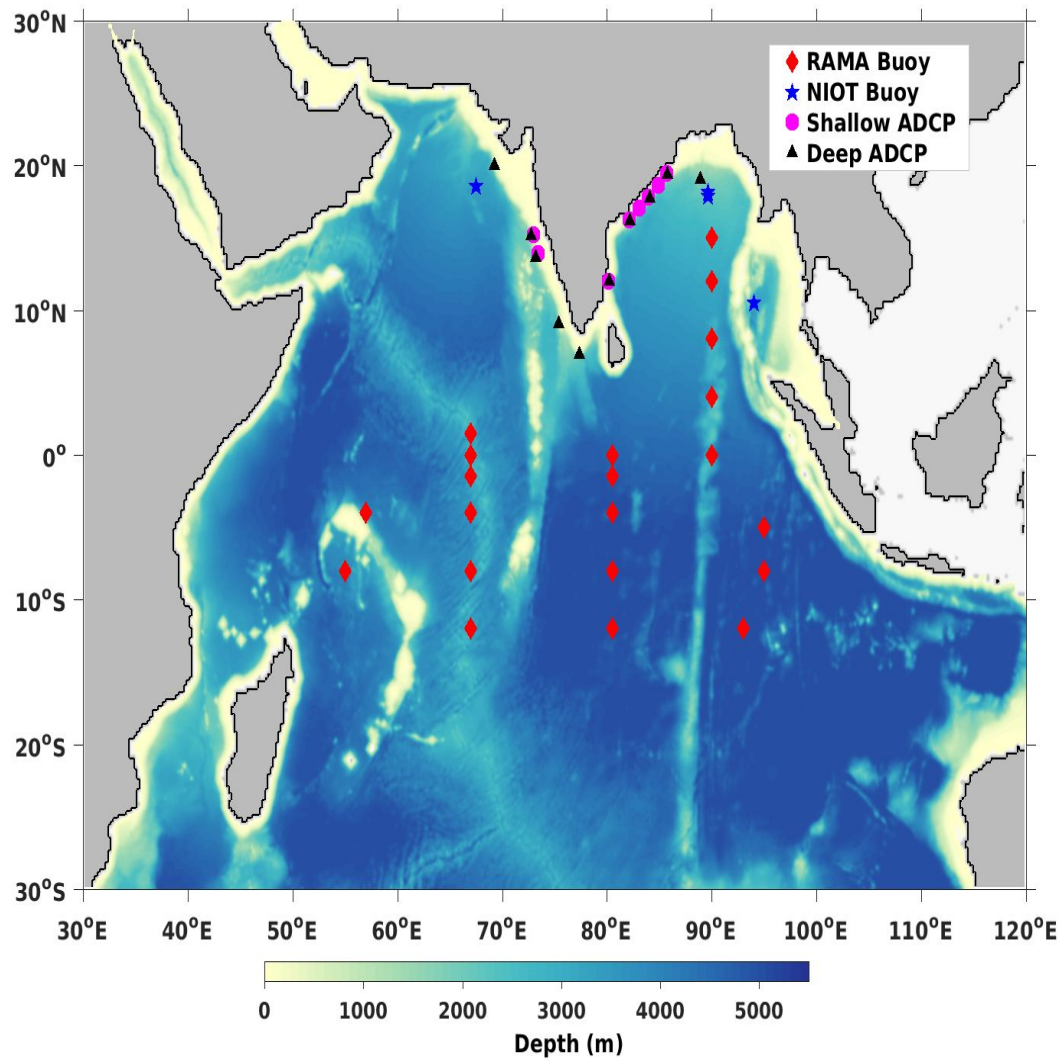
Independent Variables

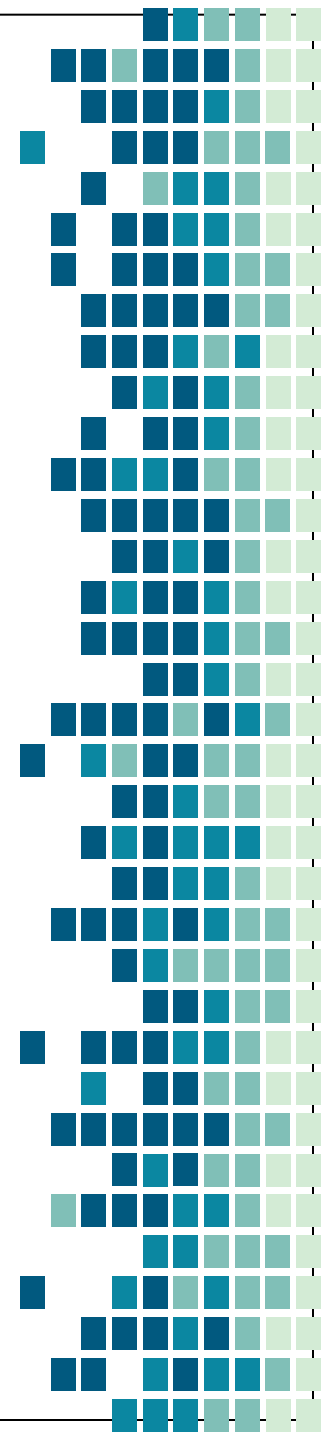
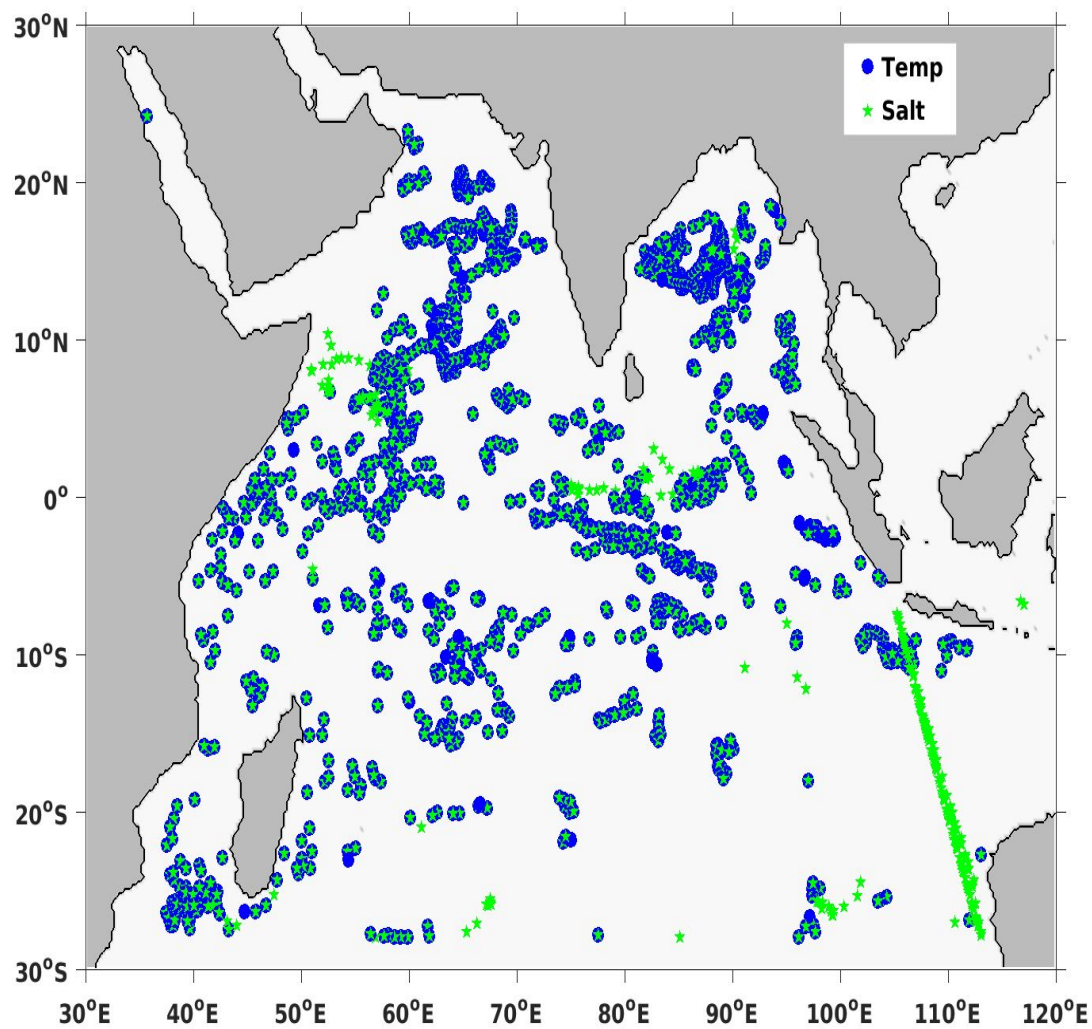
1. Sea level anomaly
2. Sea Surface salinity
3. U,V Currents

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

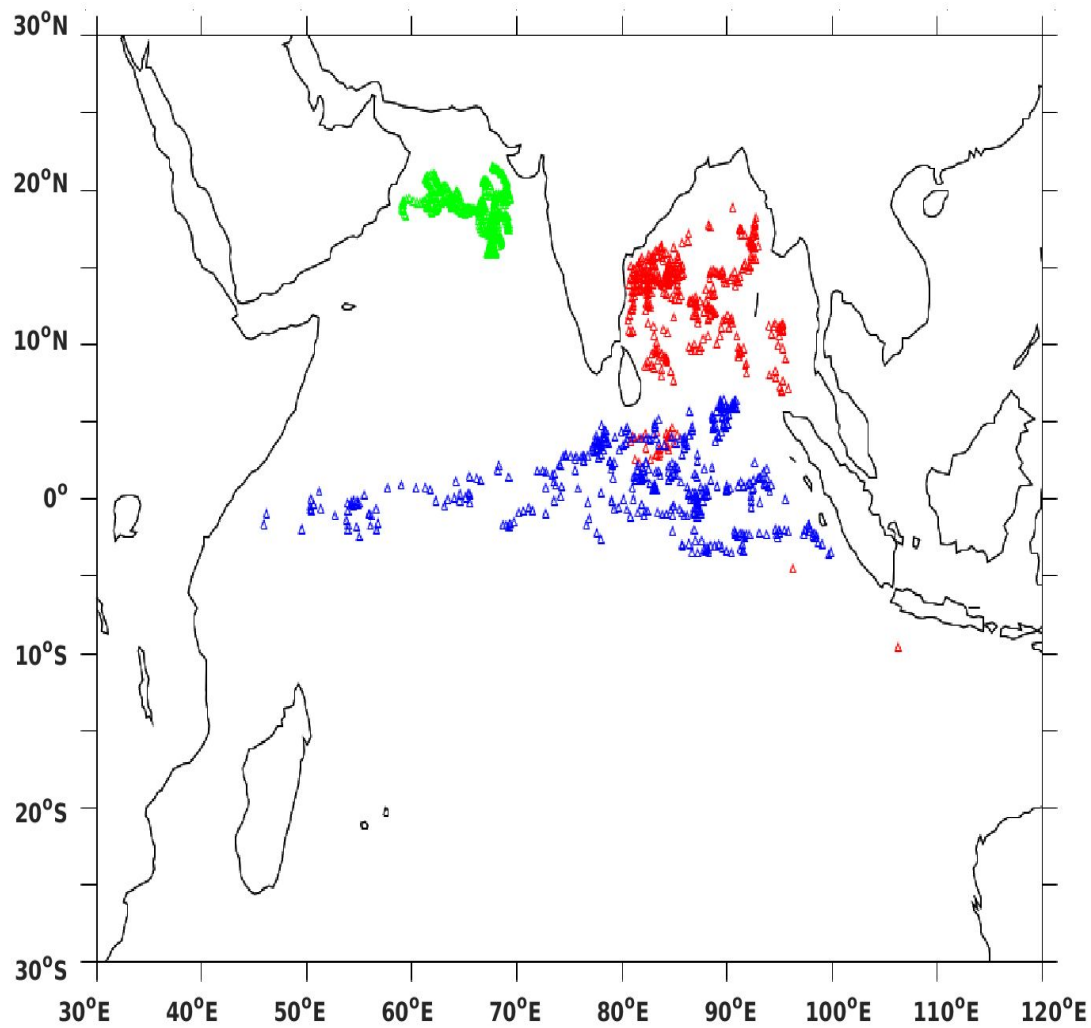


Variable	Assimilated	Validation/Comparison
SST	AMRSE satellite track data	AVHRR
SLA	-	AVISO
U V Currents	-	OSCAR, ADCP
Temperature	In-situ profiles	RAMA mooring, NIOT Buoys
Salinity	In-situ profiles	RAMA mooring, NIOT Buoys



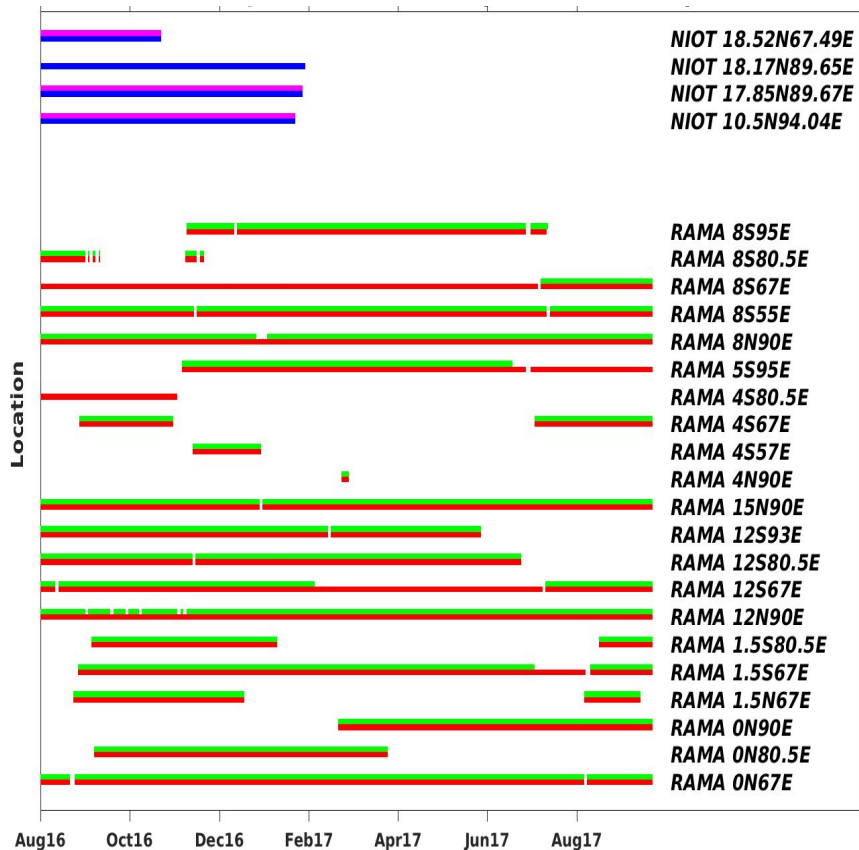


ARGO FLOATS

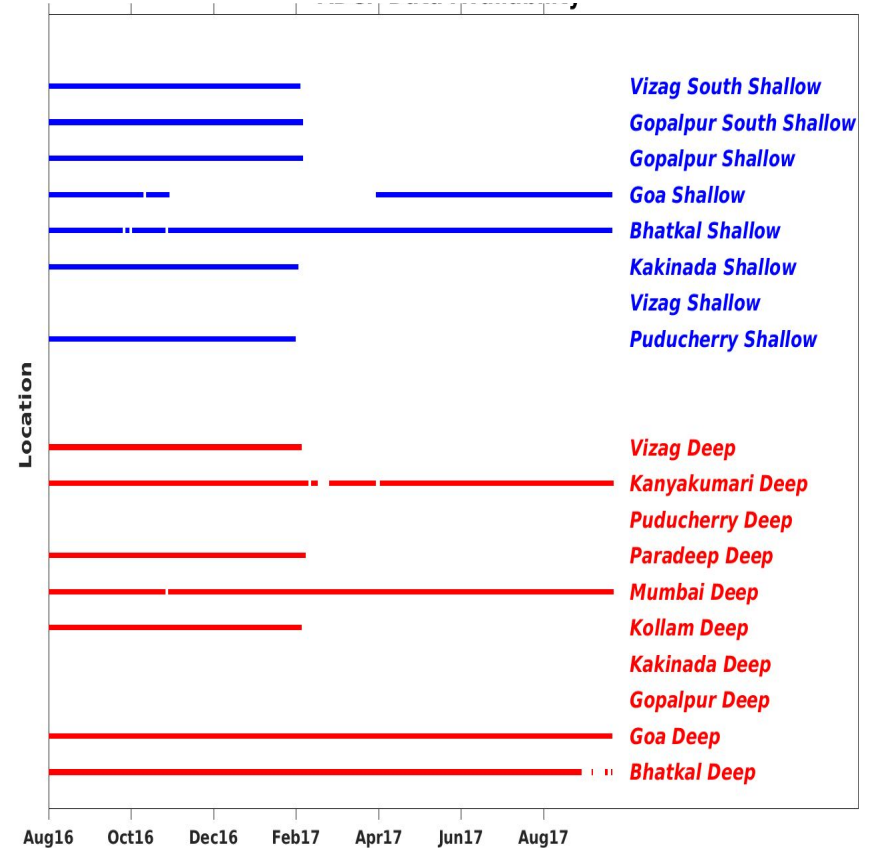


Daily pop-up of Argo floats in the Northern Indian Ocean

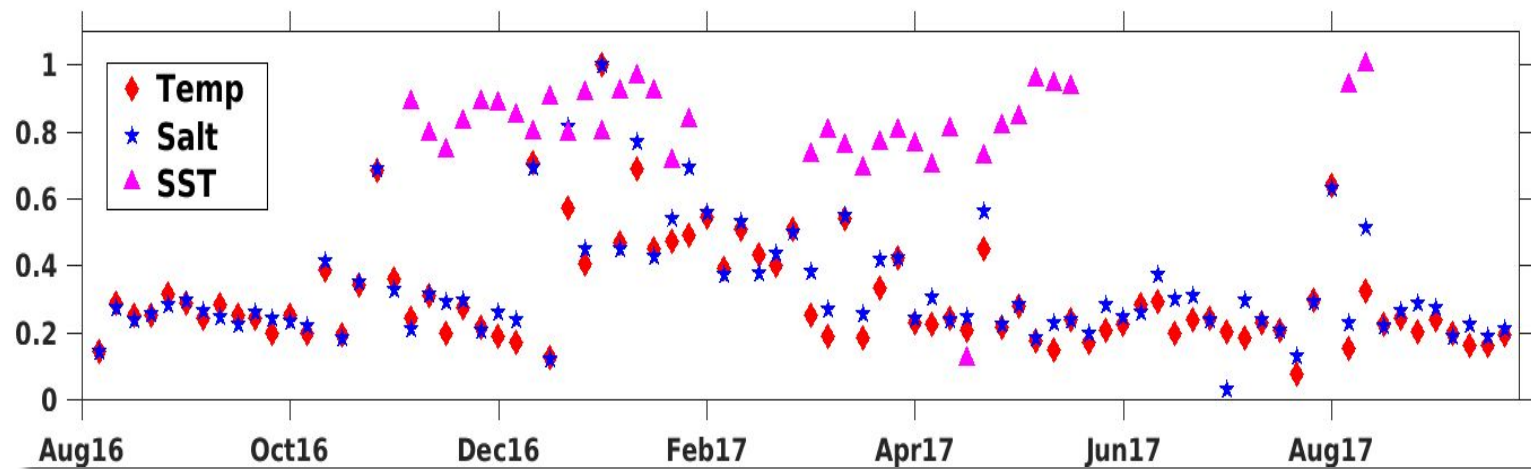
Temperature and Salinity Data availability



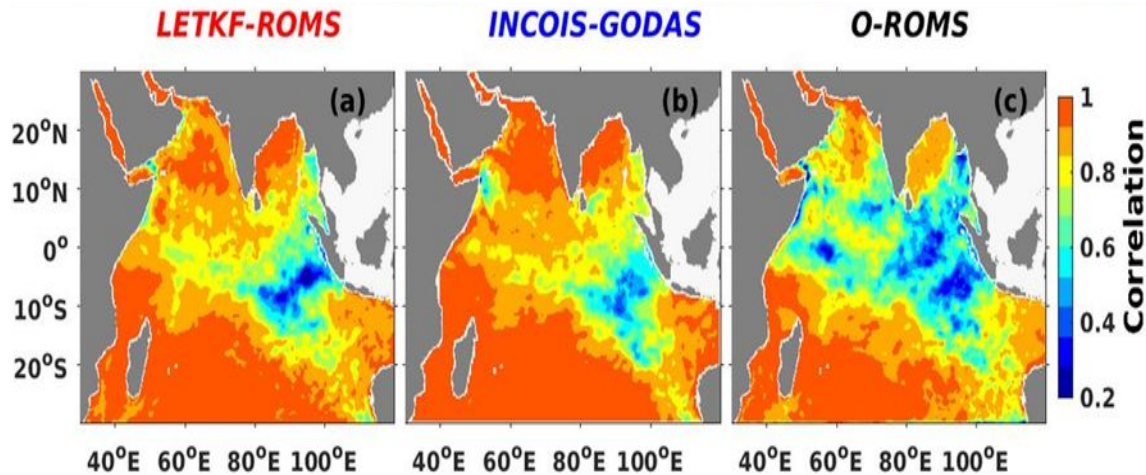
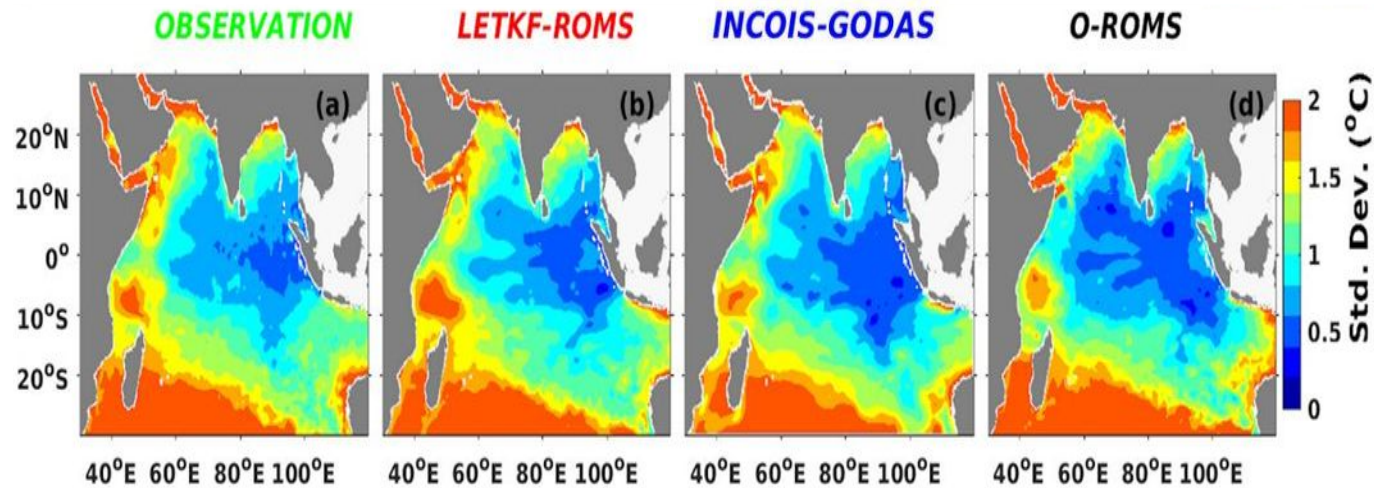
ADCP Data availability



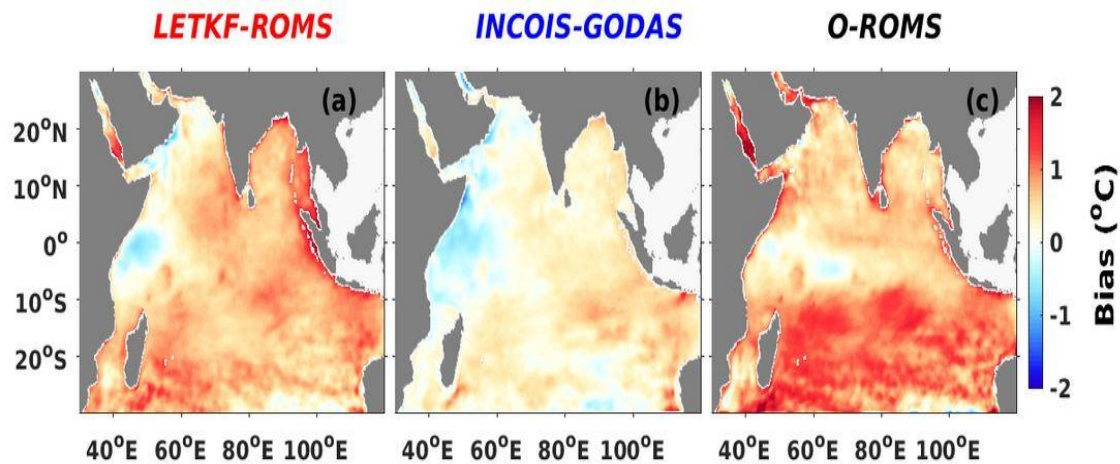
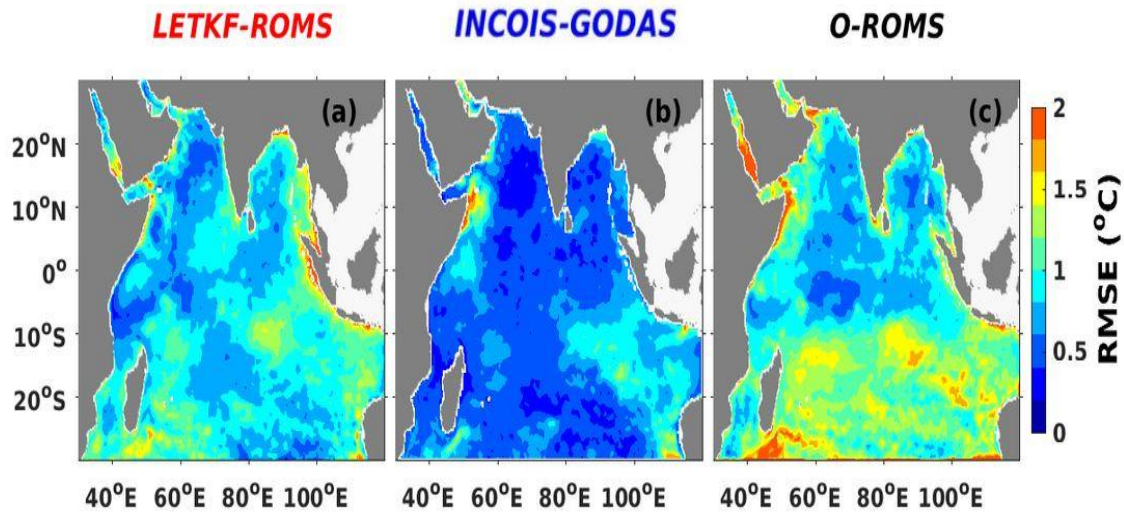
Assimilated Observations



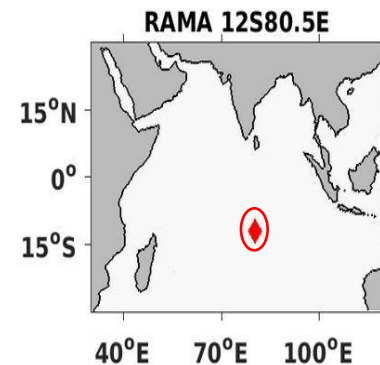
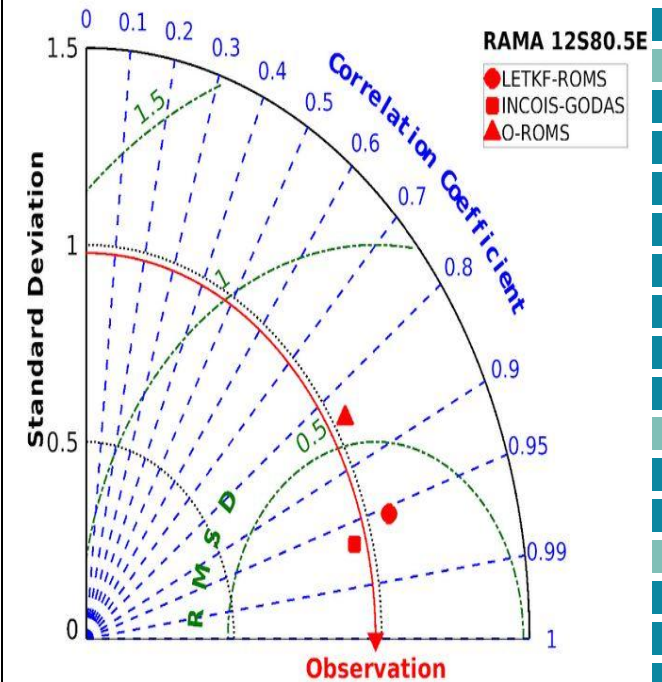
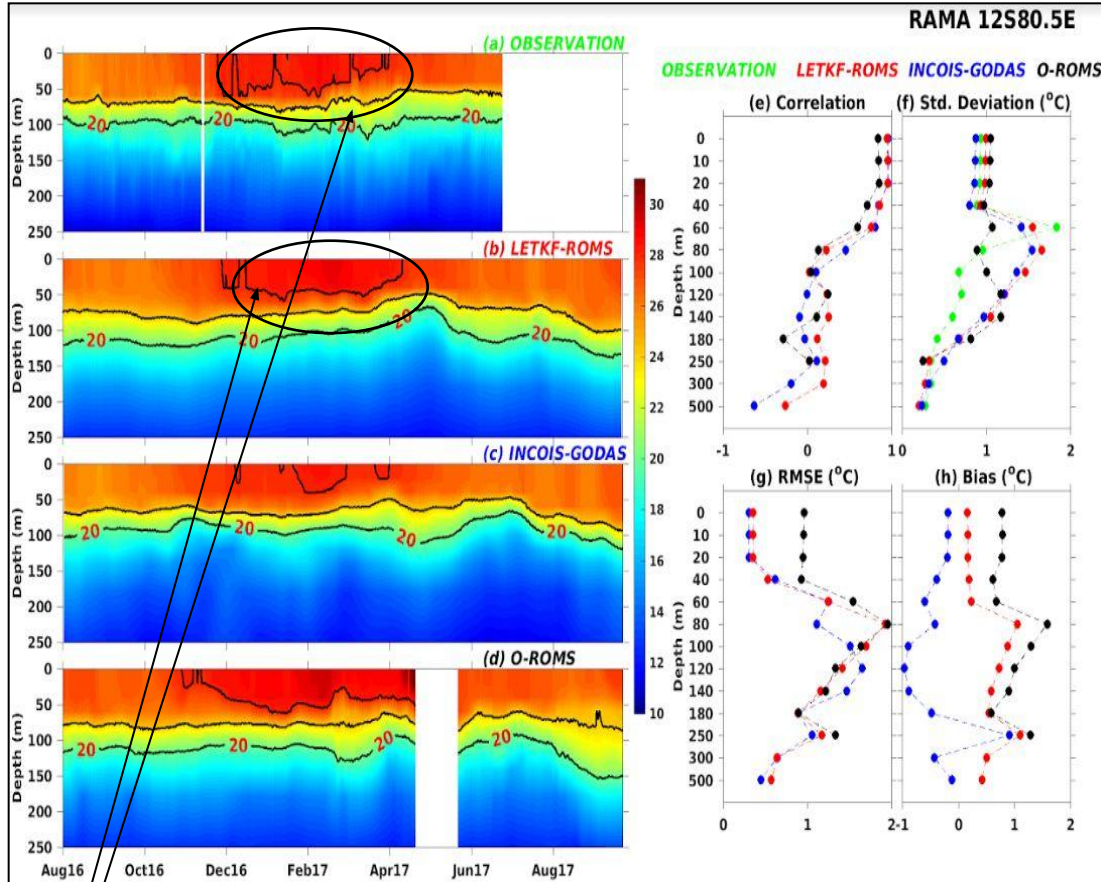
Temperature



Sea Surface Temperature



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 a best estimate.
- 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.

