Data
Assimilation
Research
Testbed

# A Randomized Dormant Ensemble Kalman Filter

## Dealing with Extreme Sampling Errors


Moha Gharamti, NCAR


18th international EnKF workshop

May 2–5, 2023

NCAR
NATIONAL CENTER FOR ATMOSPHERIC RESEARCH

NSF

# Overview

# EnKF Deficiencies

- **Sampling errors** due to small ensemble sizes $\Rightarrow$ variance underestimation, low-rank covariances, spurious correlations
- **Model biases** may lead to filter divergence
- **Nonlinearity** and **non-Gaussianity**

# EnKF Deficiencies

- **Sampling errors** due to small ensemble sizes $\Rightarrow$ variance underestimation, low-rank covariances, spurious correlations
- **Model biases** may lead to filter divergence
- **Nonlinearity** and **non-Gaussianity**

Typical remedies include:

- **Localization:** Localize the impact of the observations to nearby state variables only
- **Inflation:** Sample covariance is increased by linearly inflating each state component while preserving the mean
- **Others:** Covariance hybridization, multi-physics, nonlinear and nonGaussian filters, etc

# The Randomized Dormant EnKF (RD-EnKF)

## RD-EnKF Motivation

Given EnKF limitations: Is it possible to mitigate extreme sampling errors without the need for excessive inflation?

# The Randomized Dormant EnKF (RD-EnKF)

**RD-EnKF Motivation**

Given EnKF limitations: Is it possible to mitigate extreme sampling errors without the need for excessive inflation?

Assuming $N_e$ independent samples from the analysis pdf at time $t-1$, the forecast (prior) distribution at time $t$:

$$p\left(\mathbf{x}_t|\mathbf{y}_{1:t-1}\right) = \int_{\mathbb{R}^n} p\left(\mathbf{x}_t|\mathbf{x}_{t-1}\right) p\left(\mathbf{x}_{t-1}|\mathbf{y}_{1:t-1}\right) d\mathbf{x}_{t-1} \approx \frac{1}{N_e}\sum_{i=1}^{N_e} p\left(\mathbf{x}_t|\mathbf{x}^i_{t-1|t-1}\right),$$

the EnKF posterior then becomes:

$$\begin{aligned} p\left(\mathbf{x}_t|\mathbf{y}_{1:t}\right) &= \frac{p\left(\mathbf{y}_t|\mathbf{x}_t, \mathbf{y}_{1:t-1}\right) p\left(\mathbf{x}_t|\mathbf{y}_{1:t-1}\right)}{p\left(\mathbf{y}_t|\mathbf{y}_{1:t-1}\right)} \\ &\propto p\left(\mathbf{y}_t|\mathbf{x}_t\right) p\left(\mathbf{x}_t|\mathbf{y}_{1:t-1}\right) \equiv \frac{1}{N_e} p\left(\mathbf{y}_t|\mathbf{x}_t\right) \sum_{i=1}^{N_e} N\left(\mathbf{x}^i_{t|t-1}, \widehat{\mathbf{P}}_{t|t-1}\right). \end{aligned}$$

# RD-EnKF Update

The idea is to **randomly** break down the ensemble for each assimilated observation into 2 subsets:

- **Active:** Members that go through a regular EnKF update
- **Dormant:** Members within this subset just sit and wait

$$N_e = N_a + N_d, \quad N_d = \lfloor \alpha N_e \rceil, \quad 0 \leq \alpha \leq 1$$

$\boxed{\alpha \text{ is the } \textbf{dormancy rate}}$ ➠ $\left[ \begin{array}{ll} \alpha = 0, & \text{EnKF } \Leftrightarrow N_a = N_e \\ \alpha = 1, & \text{No DA } \Leftrightarrow N_d = N_e \end{array} \right.$

# RD-EnKF Update

The idea is to **randomly** break down the ensemble for each assimilated observation into 2 subsets:

- **Active:** Members that go through a regular EnKF update
- **Dormant:** Members within this subset just sit and wait

$$N_e = N_a + N_d, \quad N_d = \lfloor \alpha N_e \rceil, \quad 0 \le \alpha \le 1$$

$\boxed{\alpha \text{ is the } \textbf{dormancy rate}}$ $\Rightarrow$ $\left[ \begin{array}{ll} \alpha = 0, & \text{EnKF } \Leftrightarrow N_a = N_e \\ \alpha = 1, & \text{No DA } \Leftrightarrow N_d = N_e \end{array} \right.$

- After the update, the active and dormant members are aggregated
- For a large number of observations, all of the members are going to be updated but with a different observation volume
- Cheaper update than the EnKF

# RD-EnKF Illustration

**RD-EnKF**: 15 members, 2 available observations, $\alpha = 20\%$

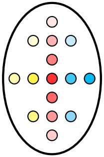
# RD-EnKF Illustration

**RD-EnKF**: 15 members, 2 available observations, $\alpha = 20\%$

obs $\#1$

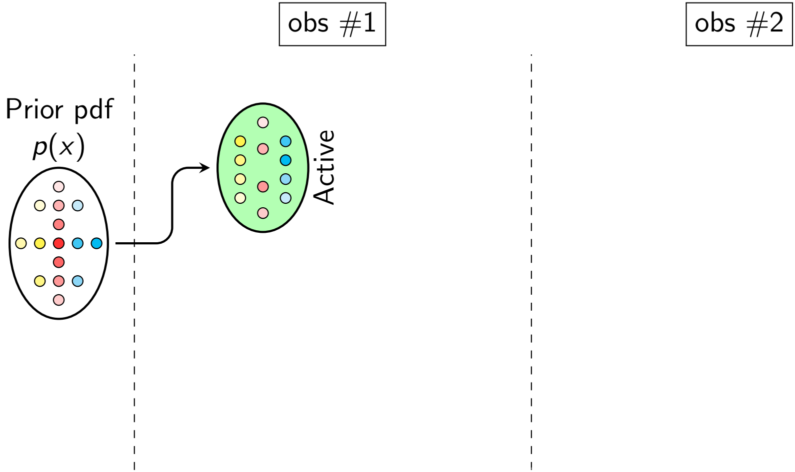
obs $\#2$

Prior pdf
$p(x)$



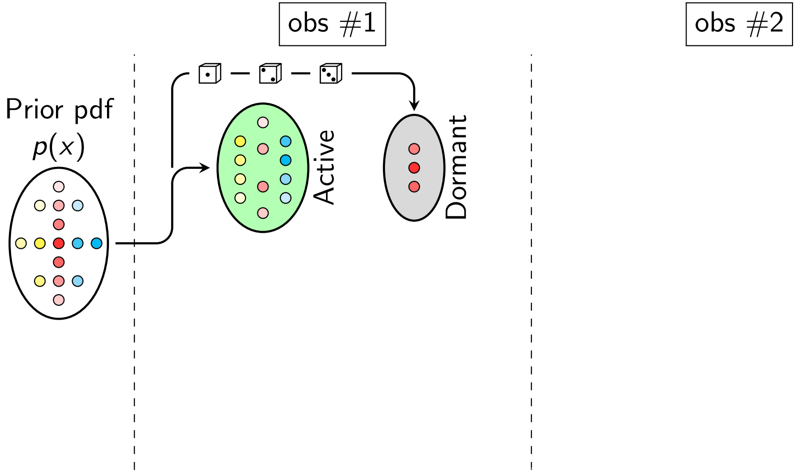
# RD-EnKF Illustration

**RD-EnKF**: 15 members, 2 available observations, $\alpha = 20\%$



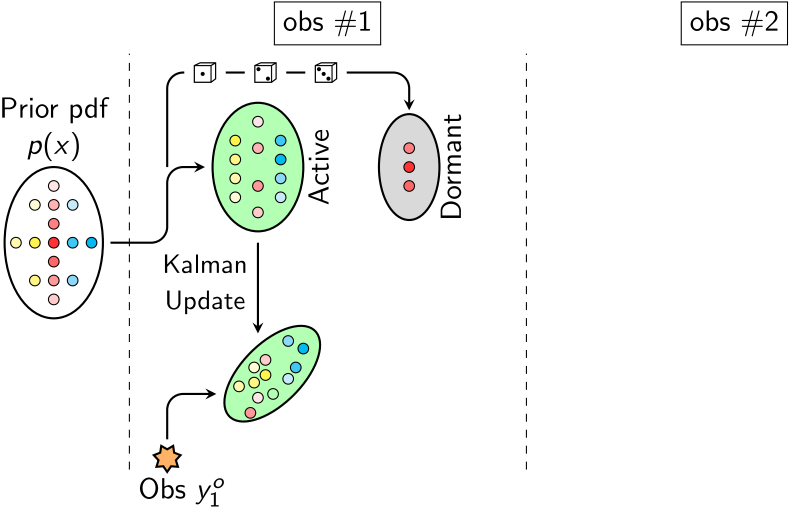
# RD-EnKF Illustration

**RD-EnKF**: 15 members, 2 available observations, $\alpha = 20\%$



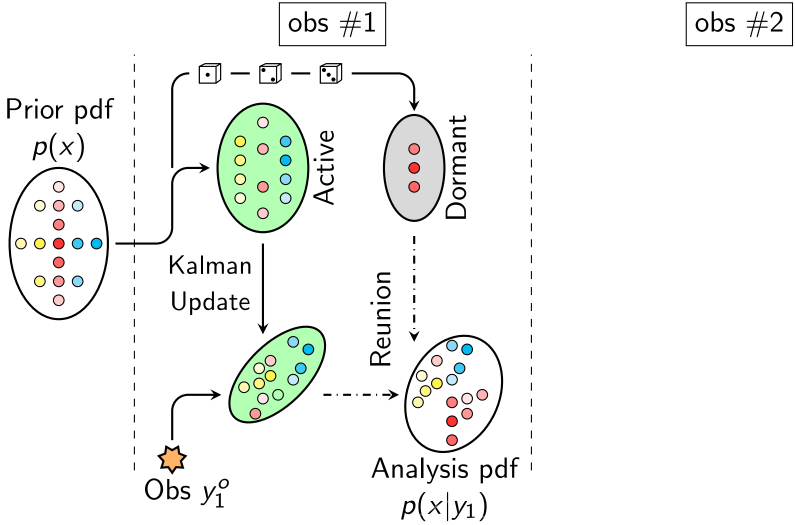
# RD-EnKF Illustration

**RD-EnKF**: 15 members, 2 available observations, $\alpha = 20\%$



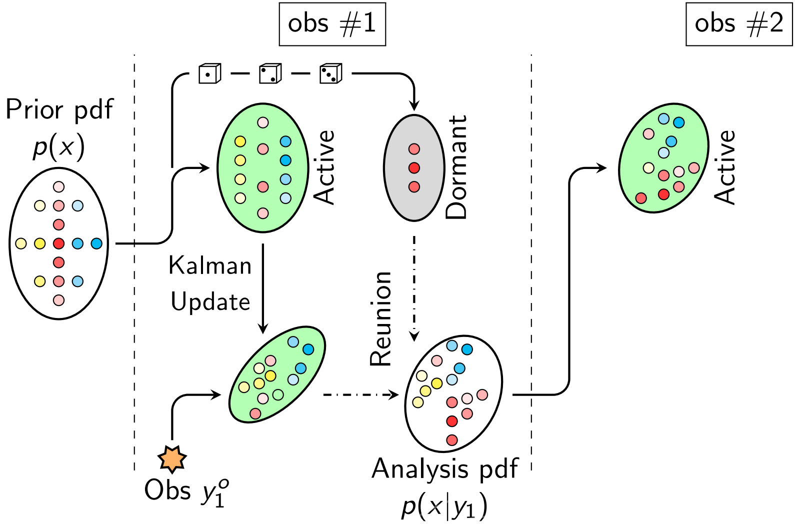
# RD-EnKF Illustration

**RD-EnKF**: 15 members, 2 available observations, $\alpha = 20\%$



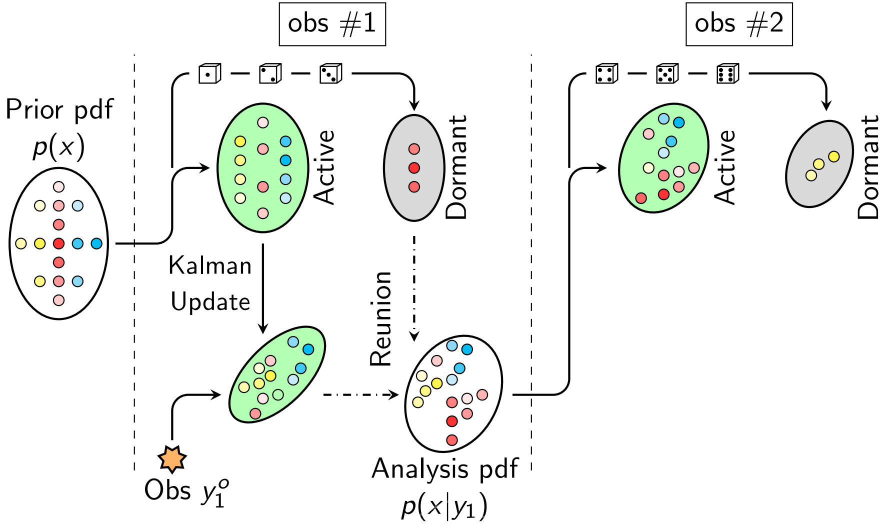
# RD-EnKF Illustration

**RD-EnKF**: 15 members, 2 available observations, $\alpha = 20\%$



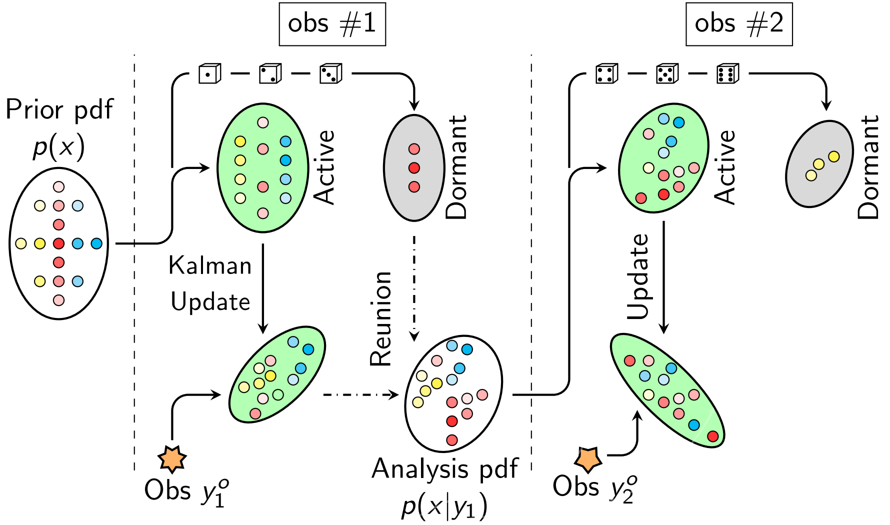
# RD-EnKF Illustration

**RD-EnKF**: 15 members, 2 available observations, $\alpha = 20\%$



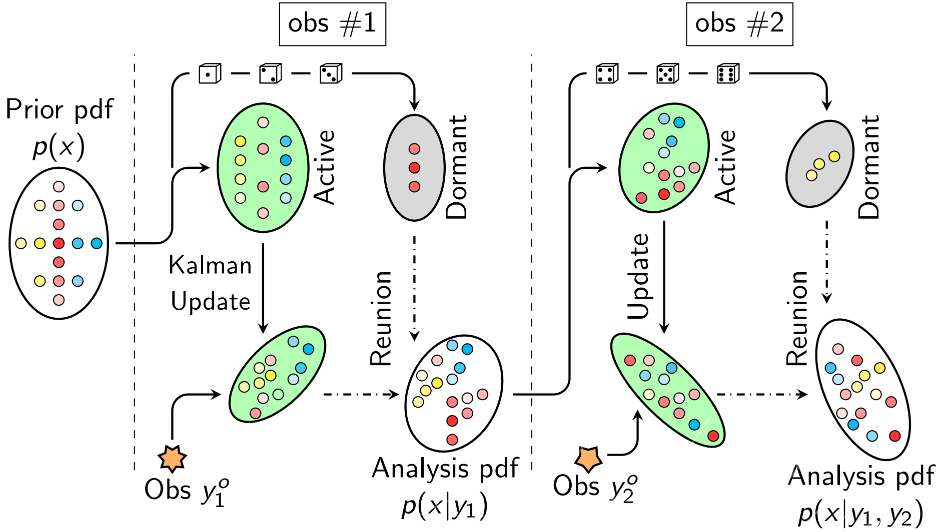
# RD-EnKF Illustration

**RD-EnKF**: 15 members, 2 available observations, $\alpha = 20\%$



# RD-EnKF Illustration

**RD-EnKF**: 15 members, 2 available observations, $\alpha = 20\%$



# RD-EnKF: Probabilistic Formulation

The EnKF samples its analysis members from this posterior density:

$$p\left(\mathbf{x}_t|\mathbf{y}_{1:t}\right) \approx p\left(\mathbf{y}_t|\mathbf{x}_t\right) p\left(\mathbf{x}_t|\mathbf{y}_{1:t-1}\right).$$

# RD-EnKF: Probabilistic Formulation

The EnKF samples its analysis members from this posterior density:

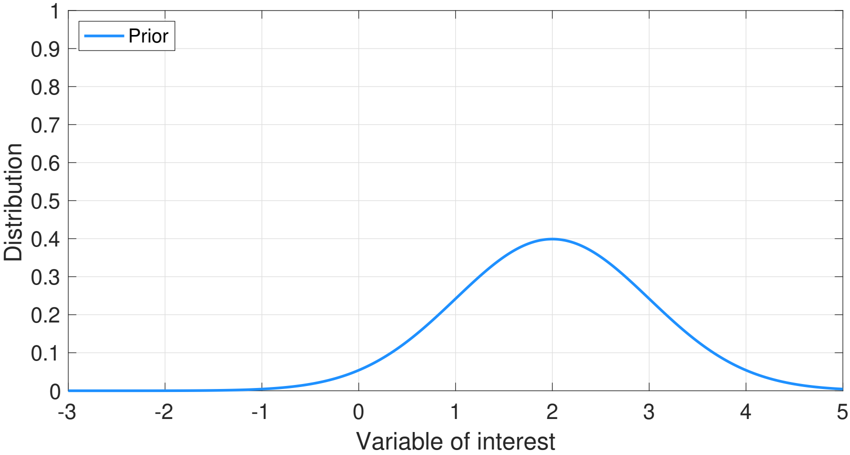
$$p\left(\mathbf{x}_t|\mathbf{y}_{1:t}\right) \approx p\left(\mathbf{y}_t|\mathbf{x}_t\right)p\left(\mathbf{x}_t|\mathbf{y}_{1:t-1}\right).$$

In contrast, the RD-EnKF samples its posterior members from a different pdf, $q\left(\mathbf{x}_t|\mathbf{y}_{1:t}\right)$ as follows:

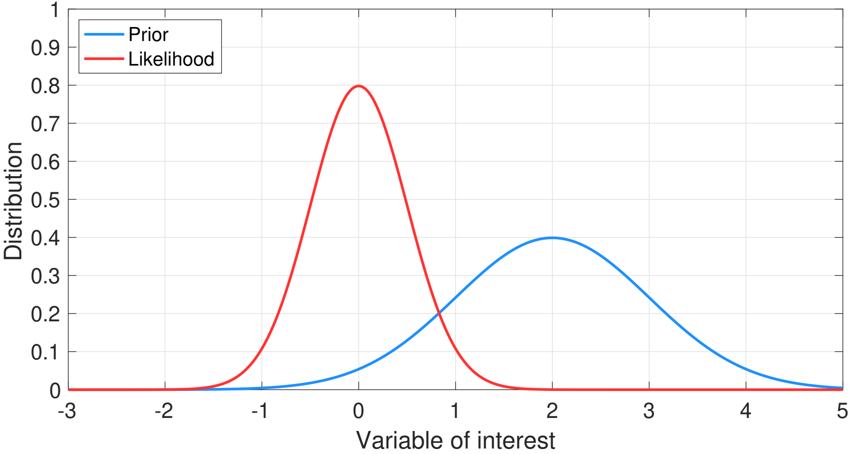
$$\begin{aligned} q\left(\mathbf{x}_t|\mathbf{y}_{1:t}\right) &= \alpha p\left(\mathbf{x}_t|\mathbf{y}_{1:t-1}\right) + (1-\alpha)p\left(\mathbf{x}_t|\mathbf{y}_{1:t}\right), \\ &\approx \alpha p\left(\mathbf{x}_t|\mathbf{y}_{1:t-1}\right) + (1-\alpha)p\left(\mathbf{y}_t|\mathbf{x}_t\right)p\left(\mathbf{x}_t|\mathbf{y}_{1:t-1}\right), \\ &\approx \underbrace{\left[\alpha + (1-\alpha)p\left(\mathbf{y}_t|\mathbf{x}_t\right)\right]}_{\text{mixture likelihood}} p\left(\mathbf{x}_t|\mathbf{y}_{1:t-1}\right), \\ &\approx \widetilde{p}\left(\mathbf{y}_t|\mathbf{x}_t\right)p\left(\mathbf{x}_t|\mathbf{y}_{1:t-1}\right). \end{aligned}$$

Both Bayesian, $p\left(\mathbf{y}_t|\mathbf{x}_t\right)$, and mixture, $\widetilde{p}\left(\mathbf{y}_t|\mathbf{x}_t\right)$, likelihoods have the same mode (assuming Gaussianity).

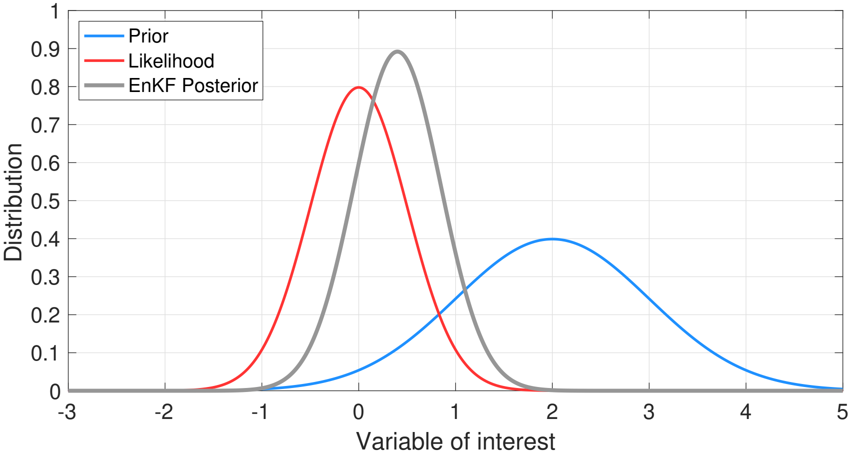
# RD-EnKF Posterior: An Example



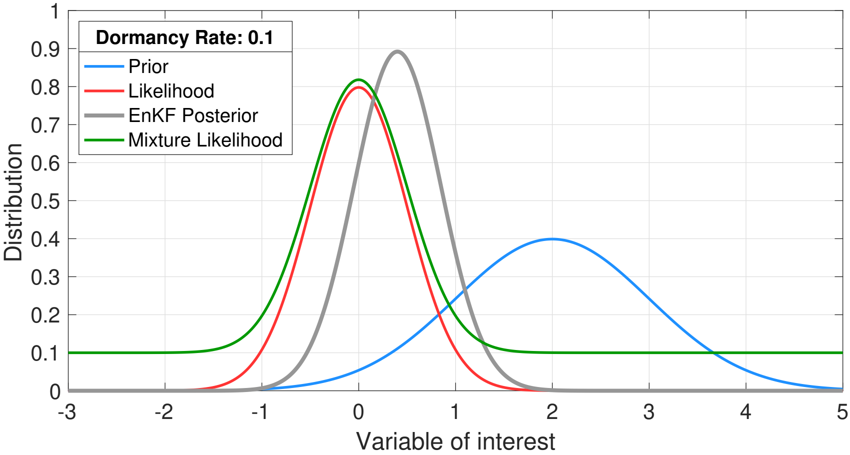
# RD-EnKF Posterior: An Example



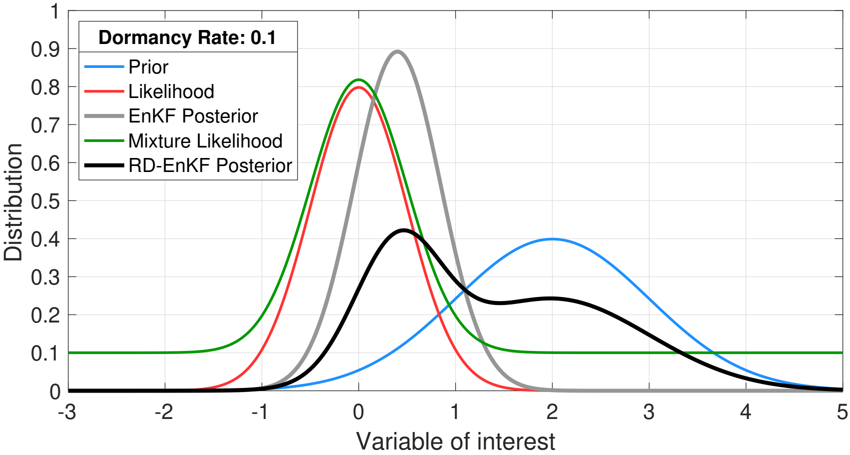
# RD-EnKF Posterior: An Example



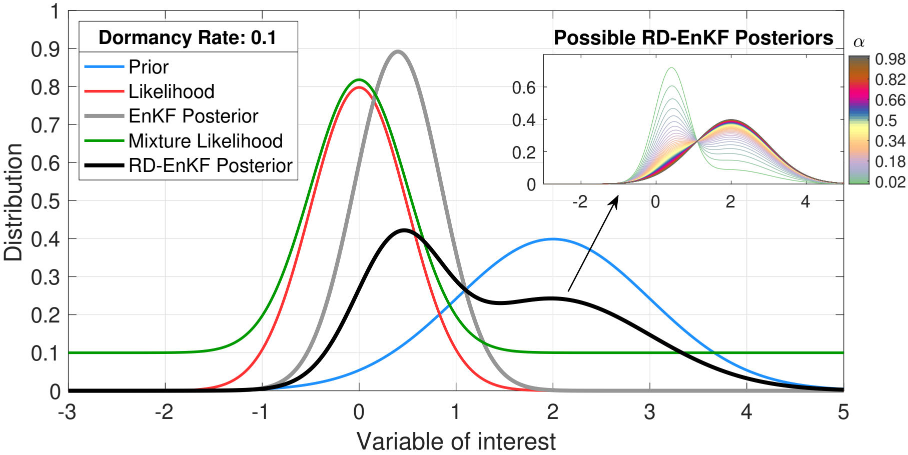
# RD-EnKF Posterior: An Example



# RD-EnKF Posterior: An Example

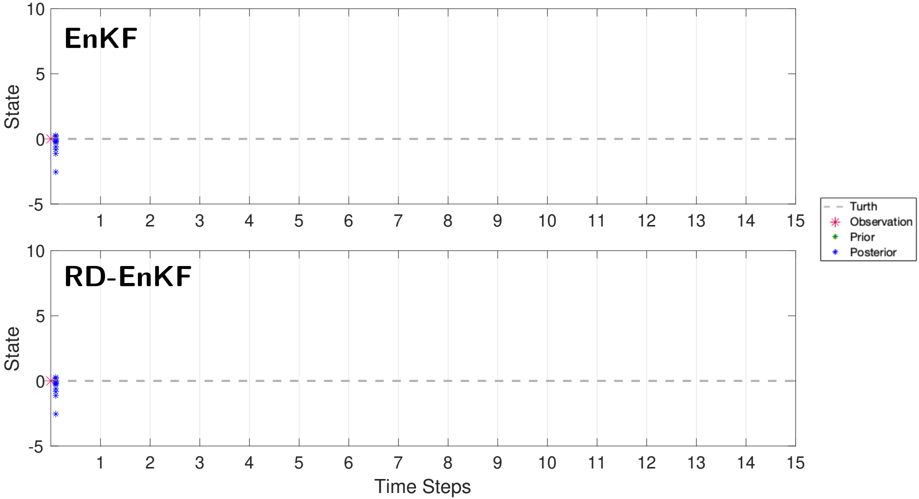


# RD-EnKF Posterior: An Example



# A Scalar Example

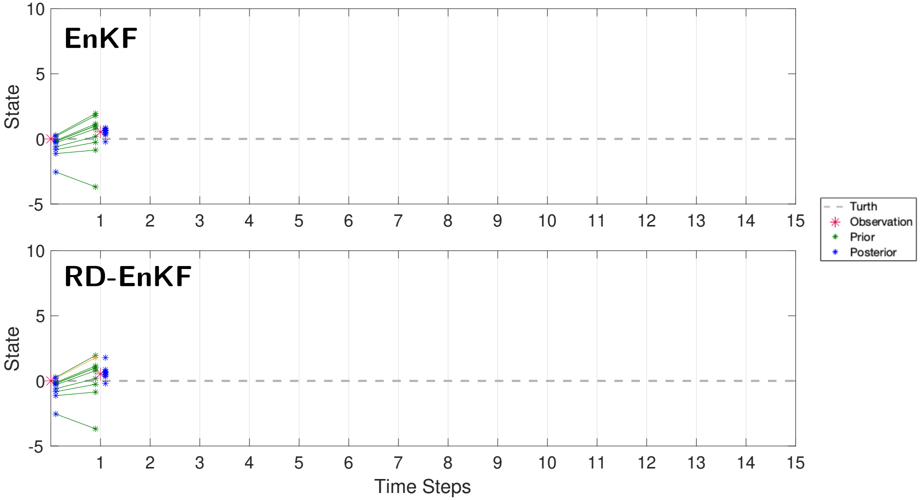
10 members, *biased* linear model, **EnKF** vs **RD-EnKF** ($\alpha = 10\%$)




NCAR
UCAR

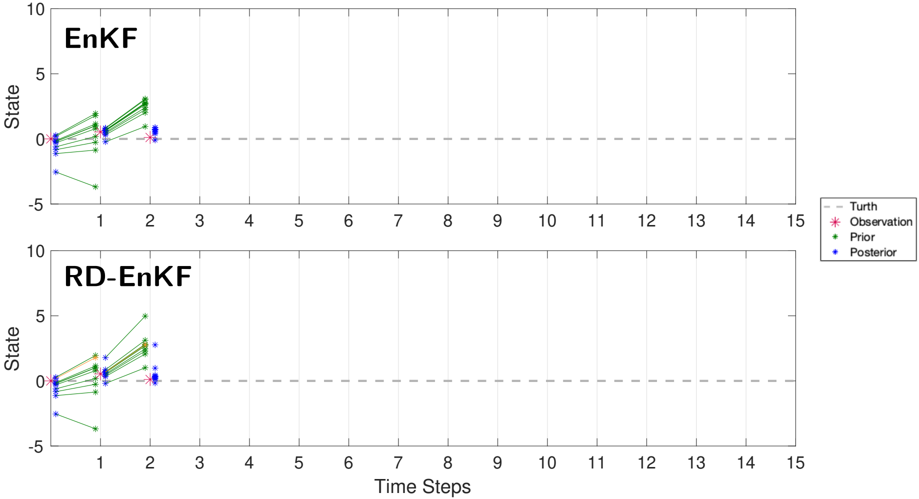
# A Scalar Example

10 members, *biased* linear model, **EnKF** vs **RD-EnKF** ($\alpha = 10\%$)



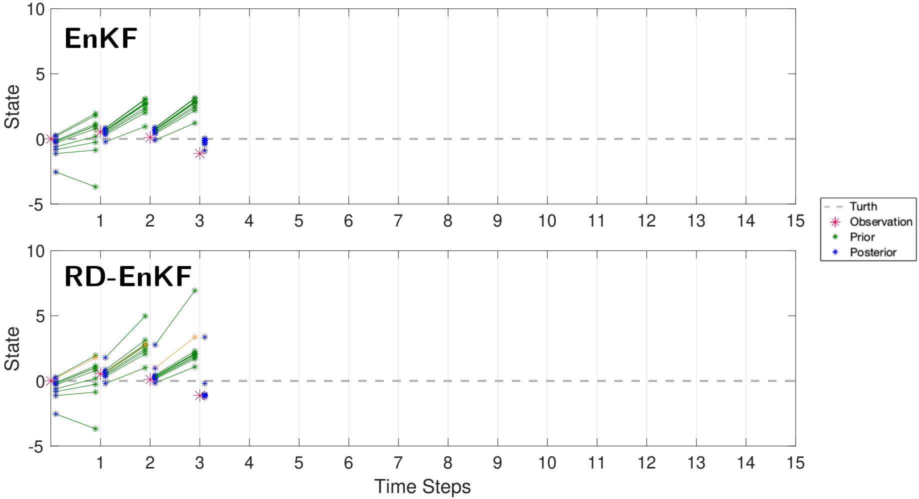
# A Scalar Example

10 members, *biased* linear model, **EnKF** vs **RD-EnKF** ($\alpha = 10\%$)



# A Scalar Example

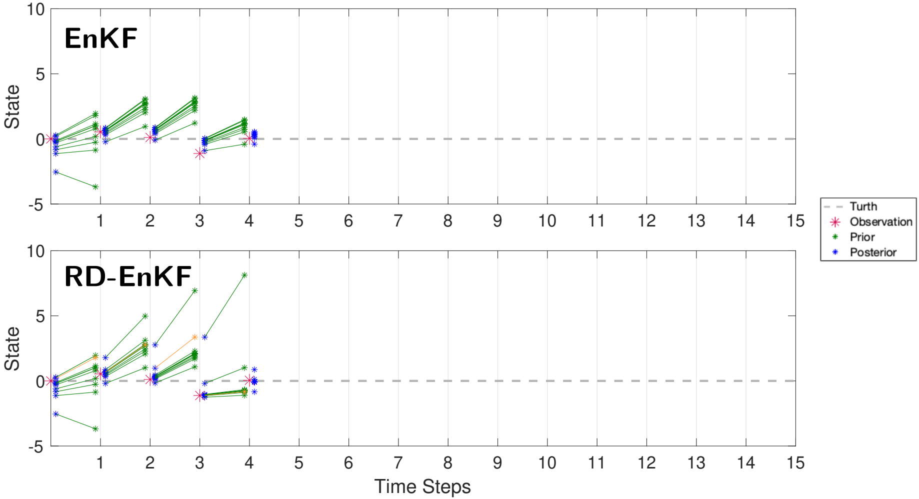
10 members, *biased* linear model, **EnKF** vs **RD-EnKF** ($\alpha = 10\%$)




NCAR
UCAR

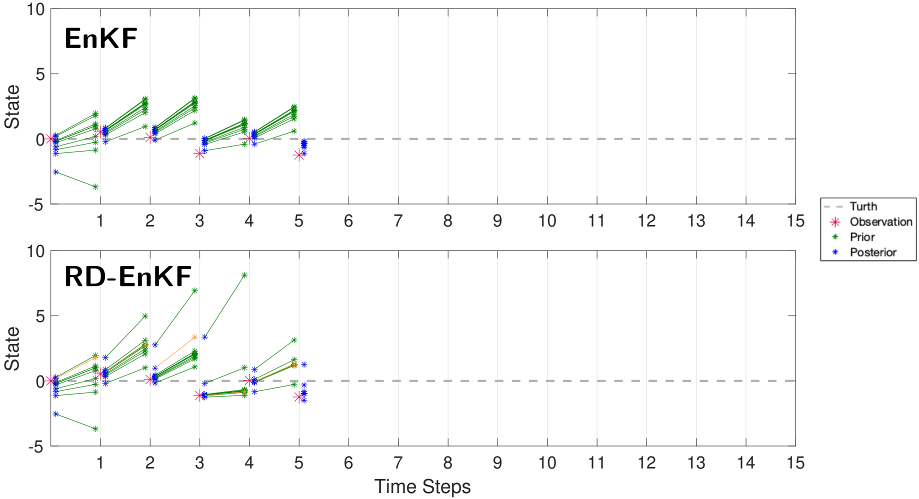
# A Scalar Example

10 members, *biased* linear model, **EnKF** vs **RD-EnKF** ($\alpha = 10\%$)



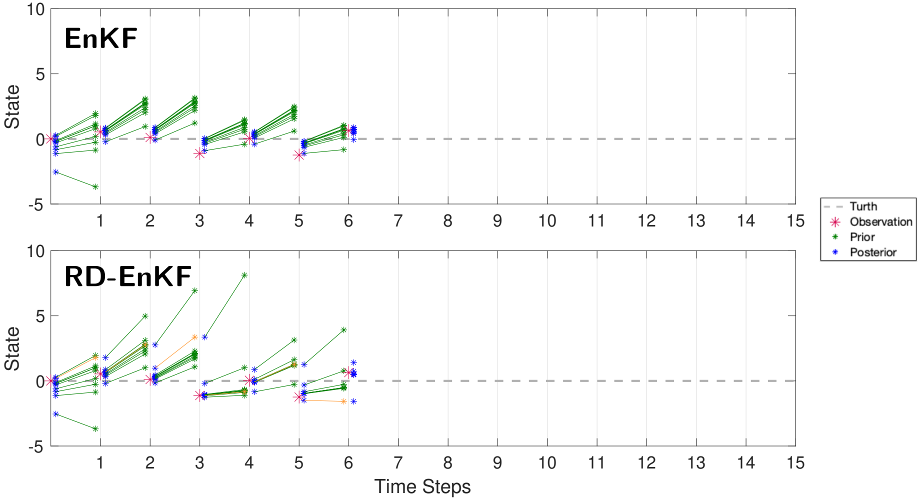
# A Scalar Example

10 members, *biased* linear model, **EnKF** vs **RD-EnKF** ($\alpha = 10\%$)



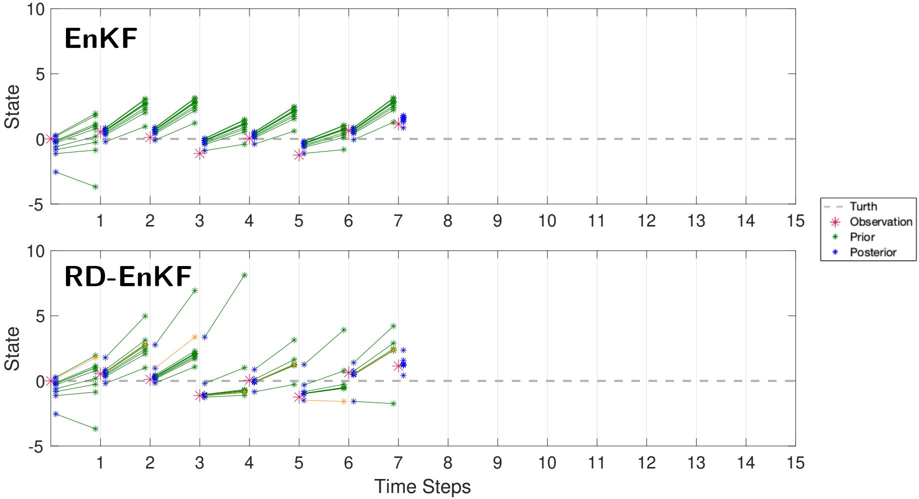
# A Scalar Example

10 members, *biased* linear model, **EnKF** vs **RD-EnKF** ($\alpha = 10\%$)



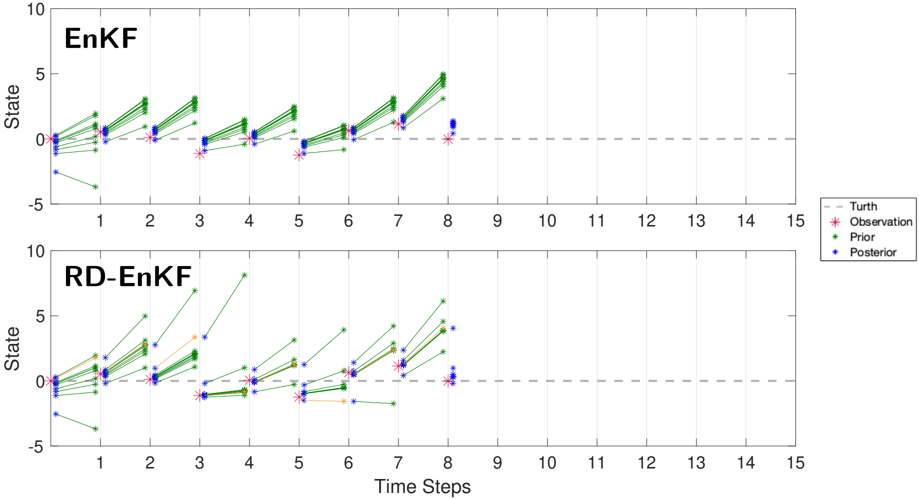
# A Scalar Example

10 members, *biased* linear model, **EnKF** vs **RD-EnKF** ($\alpha = 10\%$)



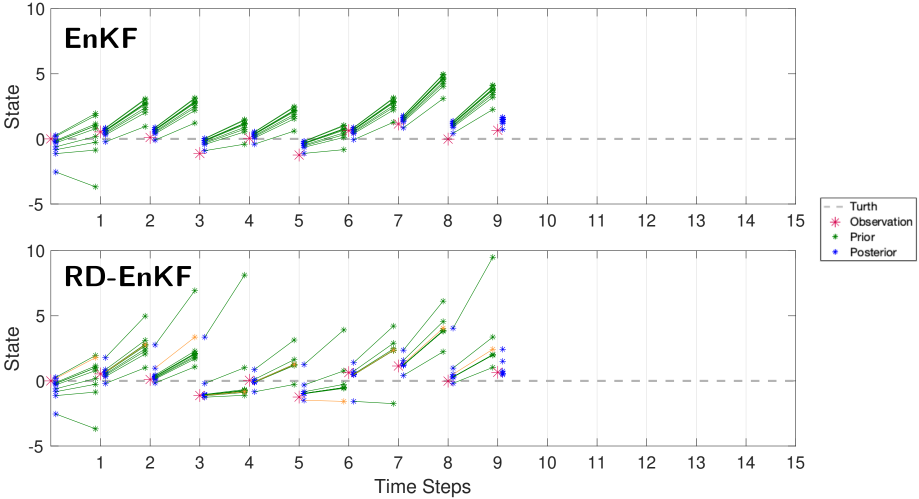
# A Scalar Example

10 members, *biased* linear model, **EnKF** vs **RD-EnKF** ($\alpha = 10\%$)



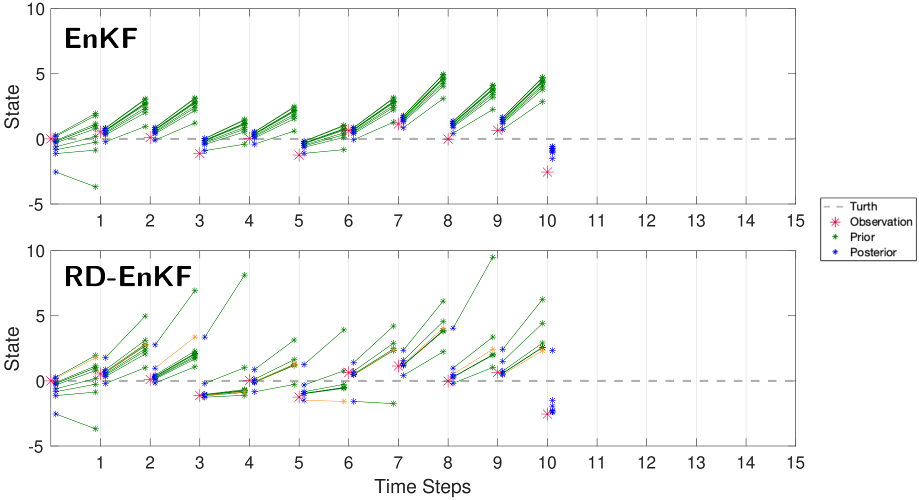
# A Scalar Example

10 members, *biased* linear model, **EnKF** vs **RD-EnKF** ($\alpha = 10\%$)



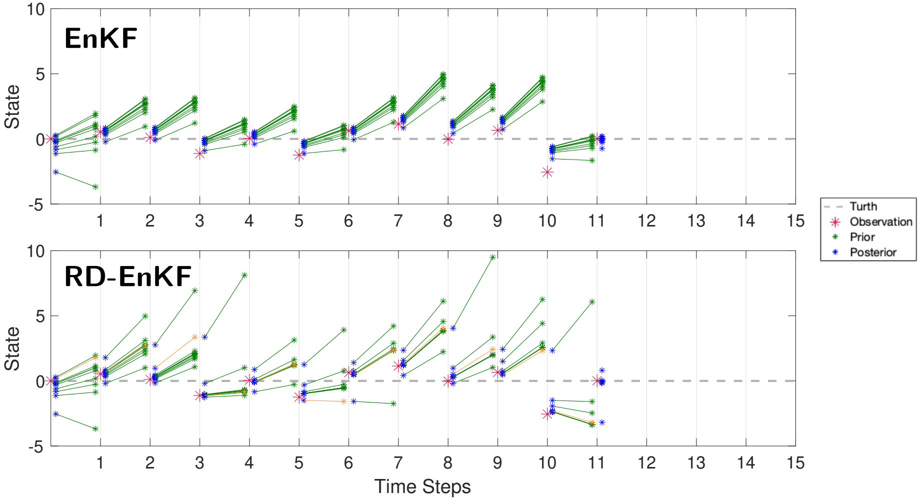
# A Scalar Example

10 members, *biased* linear model, **EnKF** vs **RD-EnKF** ($\alpha = 10\%$)



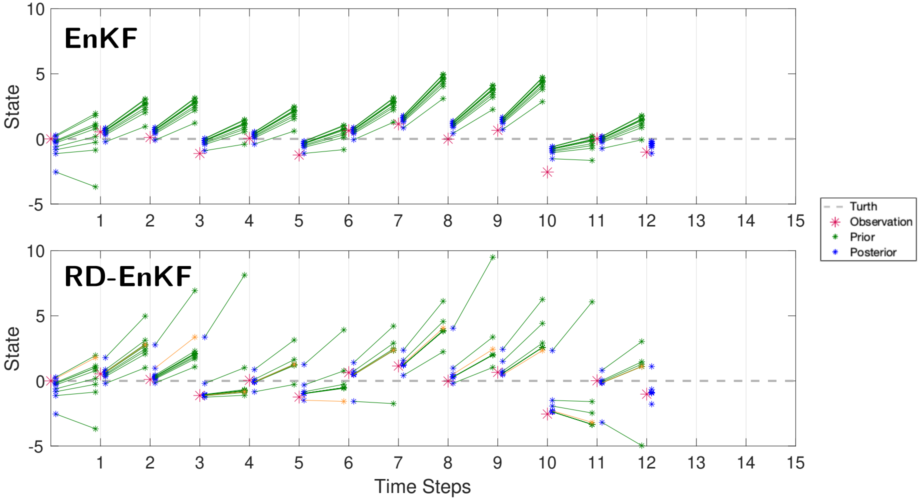
# A Scalar Example

10 members, *biased* linear model, **EnKF** vs **RD-EnKF** ($\alpha = 10\%$)



# A Scalar Example

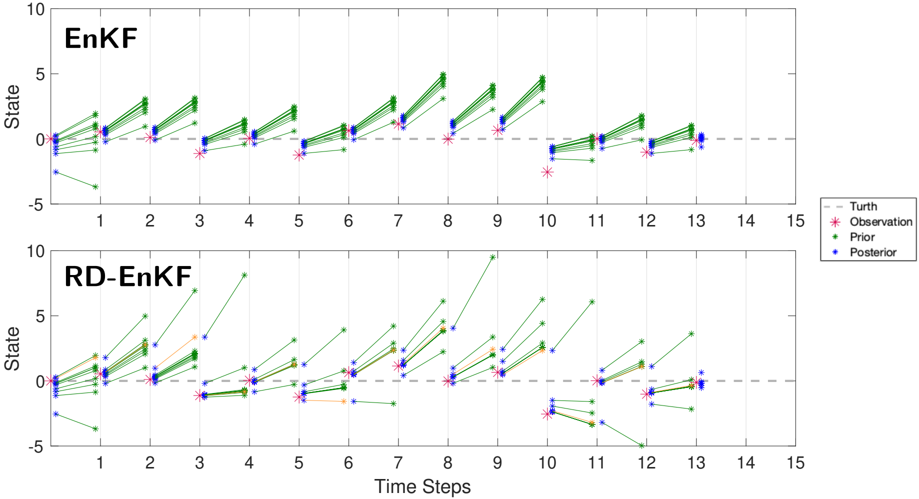
10 members, *biased* linear model, **EnKF** vs **RD-EnKF** ($\alpha = 10\%$)




NCAR
UCAR

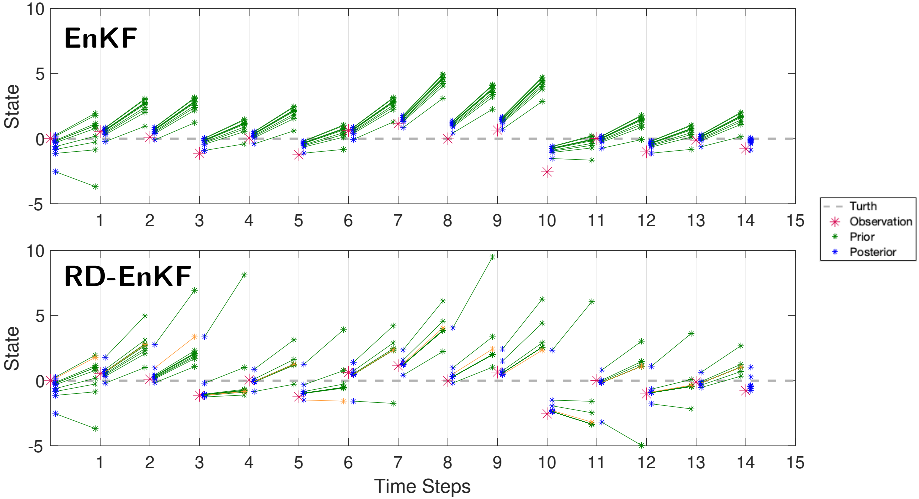
# A Scalar Example

10 members, *biased* linear model, **EnKF** vs **RD-EnKF** ($\alpha = 10\%$)



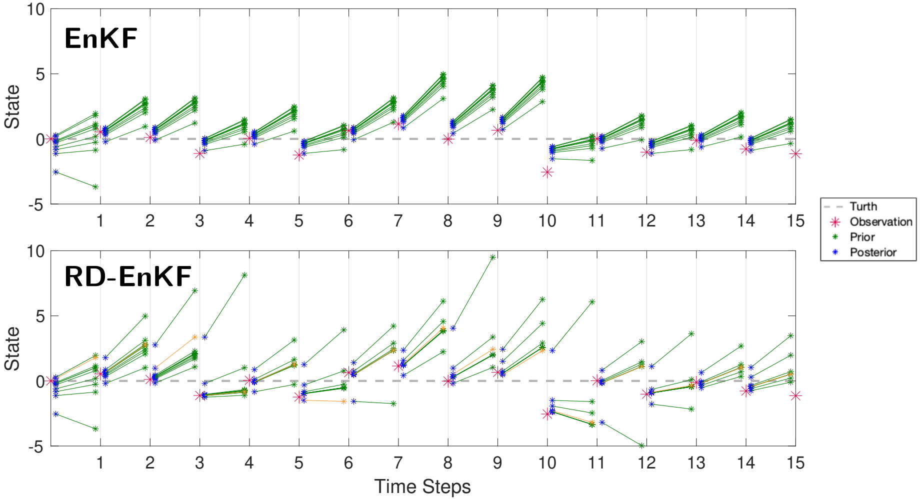
# A Scalar Example

10 members, *biased* linear model, **EnKF** vs **RD-EnKF** ($\alpha = 10\%$)



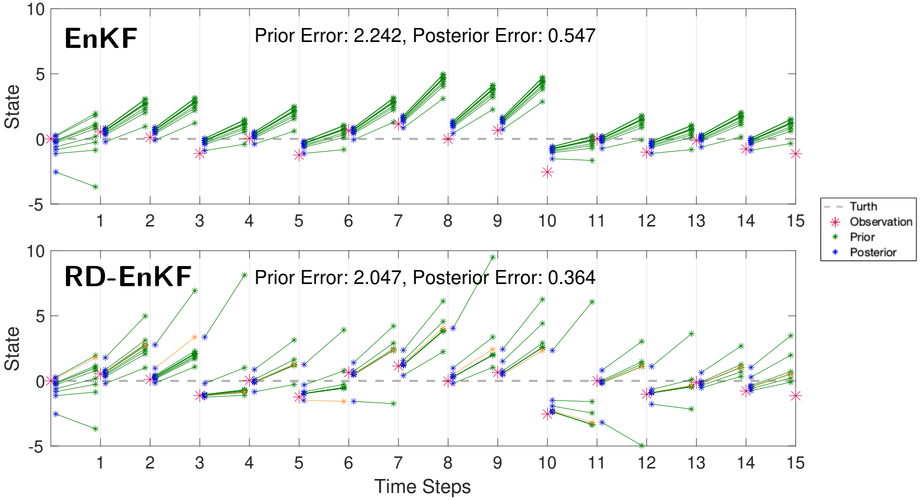
# A Scalar Example

10 members, *biased* linear model, **EnKF** vs **RD-EnKF** ($\alpha = 10\%$)



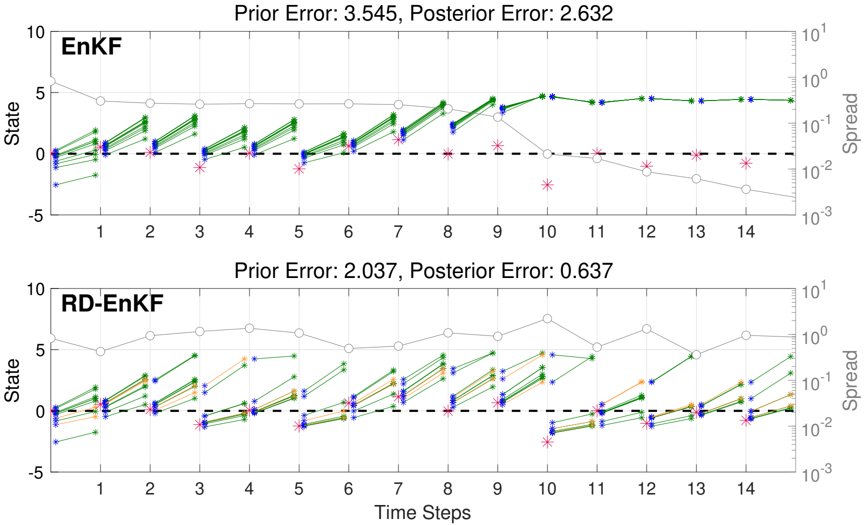
# A Scalar Example

10 members, *biased* linear model, **EnKF** vs **RD-EnKF** ($\alpha = 10\%$)



# A Scalar Example

10 members, *biased & nonlinear* model, **EnKF** vs **RD-EnKF** ($\alpha = 20\%$)



# 9-Variable Primitive Equations

Lorenz (1980) 9-variable model is a form of shallow-water equations governing hydrostatic, homogenous, incompressible, uniformly rotating flow of infinite horizontal extent.

**Variables:**
Divergence, Vorticity and Height

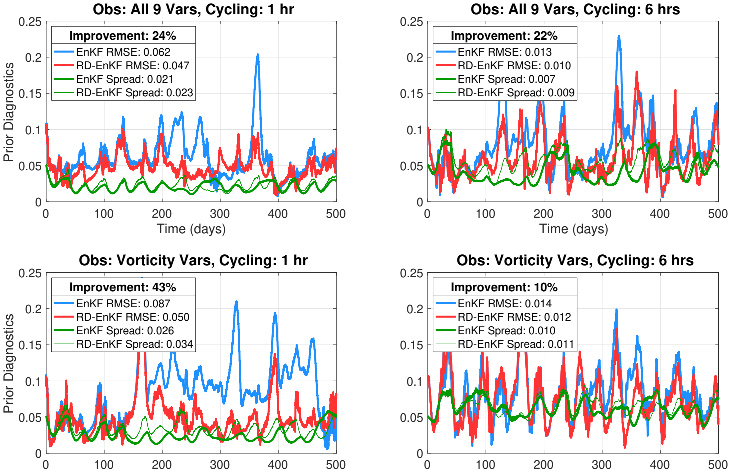
**OSSE:**

$$g_0 = \frac{gH}{f^2 L^2} = 8$$

$$\widetilde{g}_0 = 9.9$$

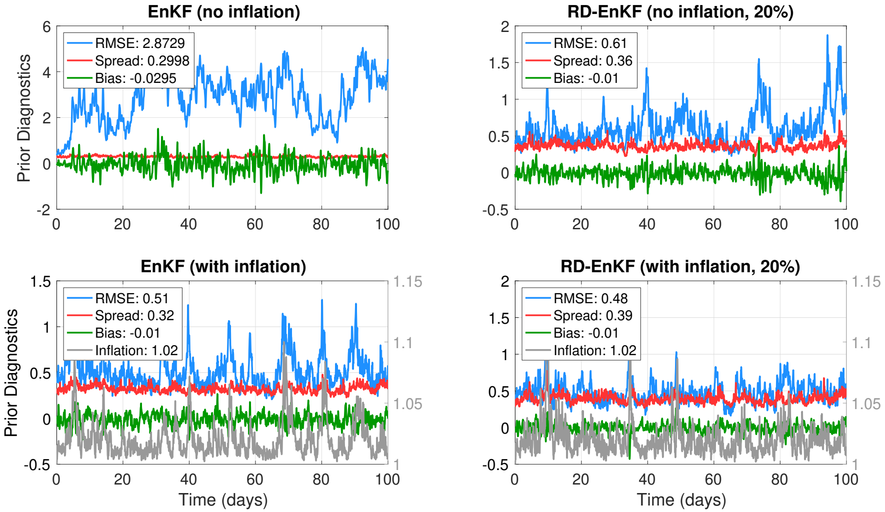
$$N_e = 20$$

$$\alpha = 15\%$$



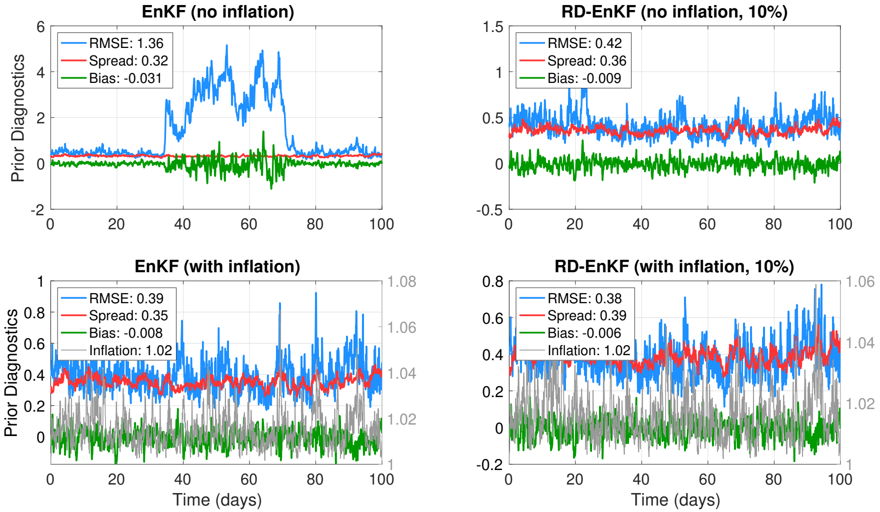
# 40-Variable Lorenz'96 Model

- No model errors, $N_e = 10$, $\alpha = 20\%$, hourly cycling, EnKF



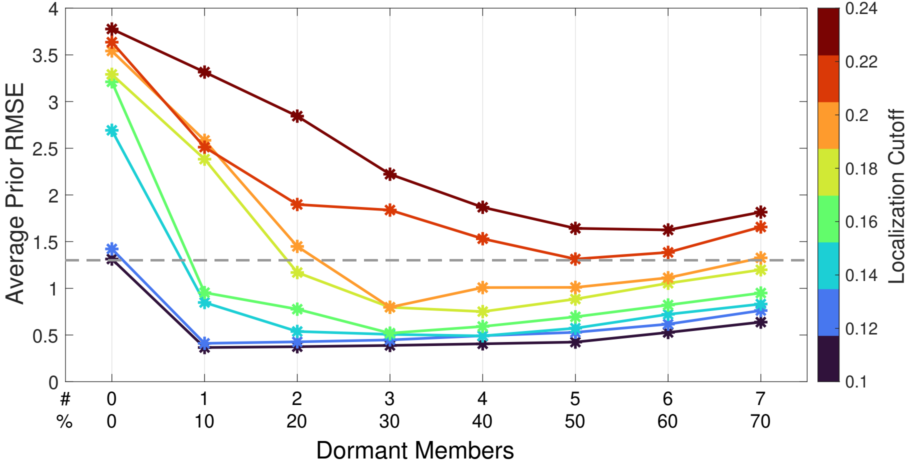
# 40-Variable Lorenz'96 Model

- No model errors, $N_e = 10$, $\alpha = 20\%$, hourly cycling, EAKF



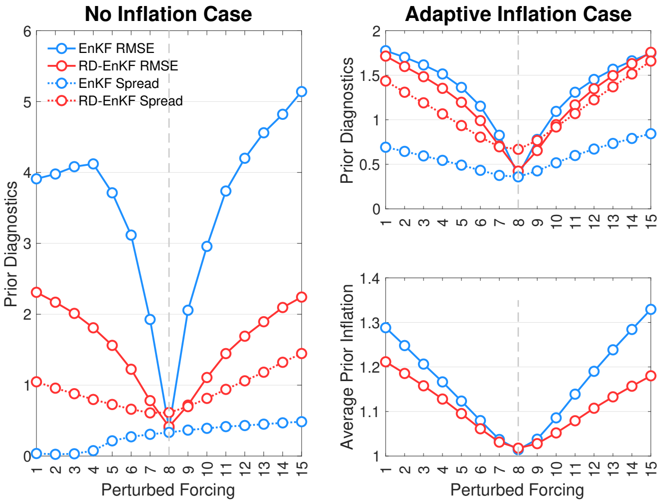
# 40-Variable Lorenz'96 Model

- No model errors, $N_e = 10$, Localization Sensitivity

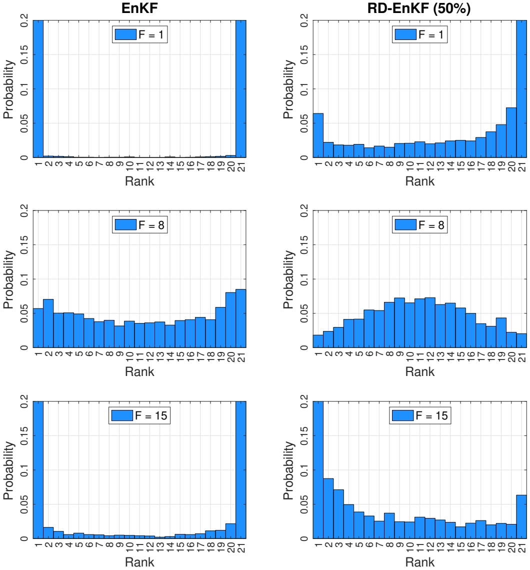


# 40-Variable Lorenz'96 Model

- With model errors, $N_e = 20$, $\alpha = 30\%$, hourly cycling

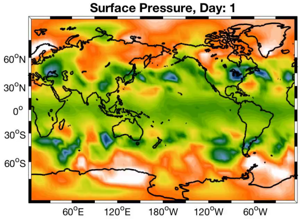


# 40-Variable Lorenz'96 Model



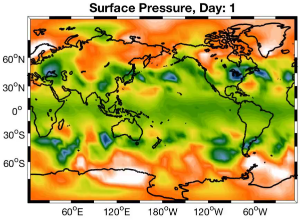
# Idealized Global Atmospheric GCM

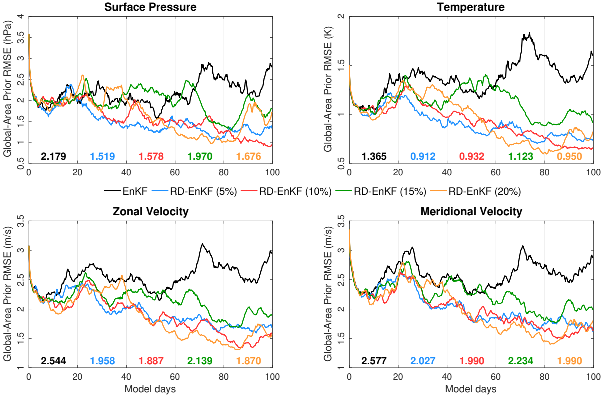
- Dry dynamical core of GFDL's AM2 model
- 30 lat $\times$ 60 lon $\times$ 5 lev $\approx 3 \times 10^4$
- Prognostic variables: **PS**, **T**, **U**, **V**
- 20 ensemble members



# Idealized Global Atmospheric GCM

- Dry dynamical core of GFDL's AM2 model
- 30 lat $\times$ 60 lon $\times$ 5 lev $\approx 3 \times 10^4$
- Prognostic variables: **PS**, **T**, **U**, **V**
- 20 ensemble members

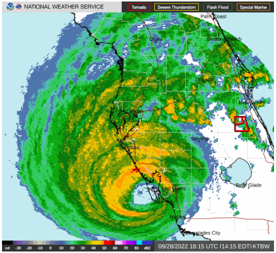



- 300 PS observations
- 6-hr cycling
- Best accuracy for $\alpha = 5, 10\%$

# Flood and Streamflow Prediction: Hurricane Ian

- Category 4 hurricane
- Landfall on Sep. 28$^{th}$, 2022
- 5$^{th}$ strongest on US record
- Wind gusts: 215 mph
- Precip exceeded 20 inches
- Death toll: 148
- Damages $\sim$ \$50 bn

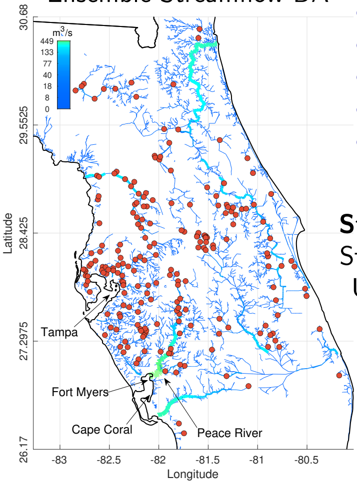



# Flood and Streamflow Prediction: Hurricane Ian

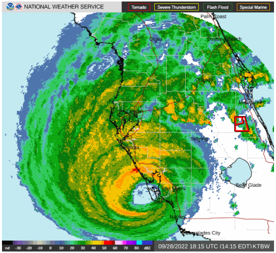
**HydroDART**

[El Gharamti et al., 2021]

Ensemble Streamflow DA




- Category 4 hurricane
- Landfall on Sep. 28$^{th}$, 2022
- 5$^{th}$ strongest on US record
- Wind gusts: 215 mph
- Precip exceeded 20 inches
- Death toll: 148
- Damages ∼ $50 bn

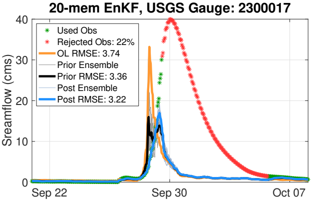


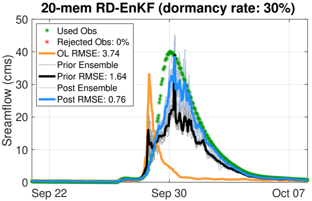

**Study Domain**

Streamflow and USGS gauges

WRF-Hydro®

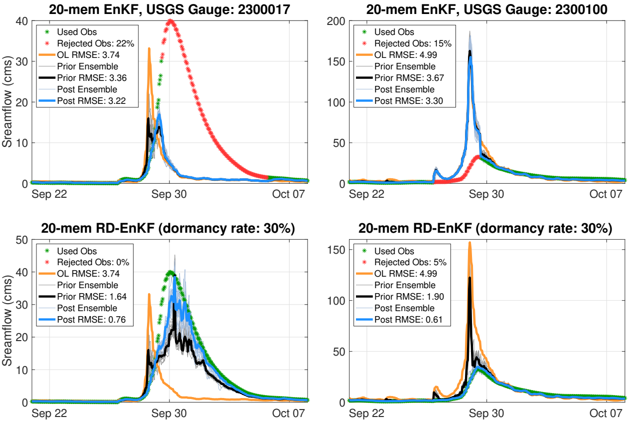
# Flood and Streamflow Prediction: Hurricane Ian



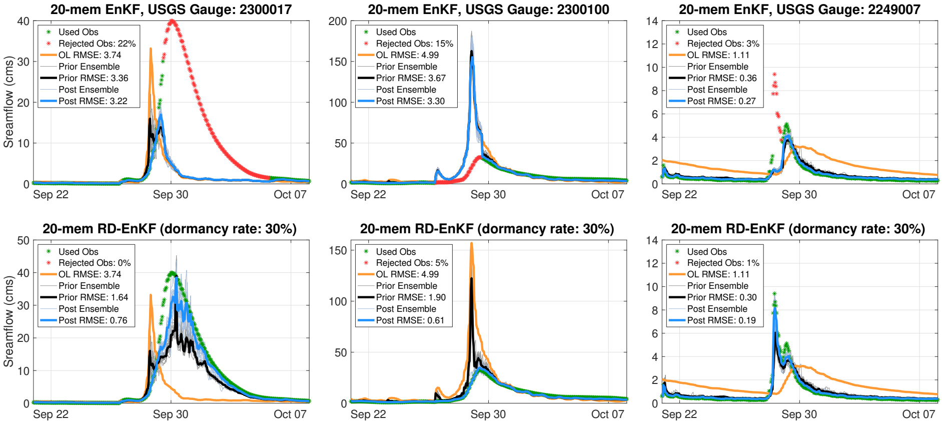

- **Variables:** Streamflow (18,190 links), groundwater bucket
- **Observations:** 1-month DA period, hourly streamflow data
- **DA Configuration:** Multiphysics, Adaptive inflation, ATS Localization

# Flood and Streamflow Prediction: Hurricane Ian



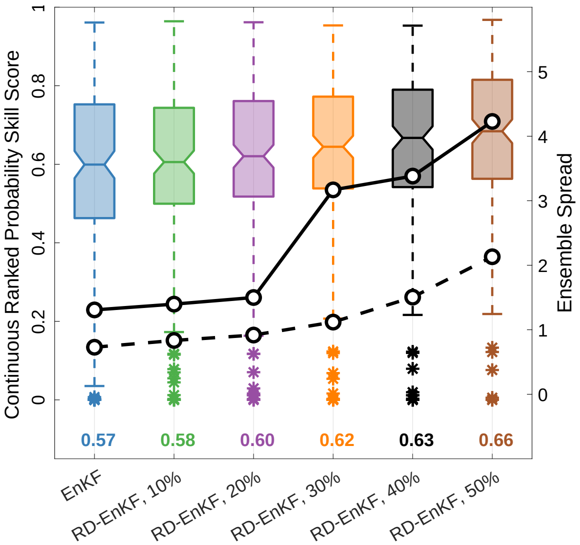

- **Variables:** Streamflow (18,190 links), groundwater bucket
- **Observations:** 1-month DA period, hourly streamflow data
- **DA Configuration:** Multiphysics, Adaptive inflation, ATS Localization

# Flood and Streamflow Prediction: Hurricane Ian




- **Variables:** Streamflow (18,190 links), groundwater bucket
- **Observations:** 1-month DA period, hourly streamflow data
- **DA Configuration:** Multiphysics, Adaptive inflation, ATS Localization

# Flood and Streamflow Prediction: Hurricane Ian



# Conclusion

- A new method for tackling sampling errors is proposed
- The randomized dormant EnKF is specifically designed for systems with limited uncertainty growth and extreme sampling error scenarios

# Conclusion

- A new method for tackling sampling errors is proposed
- The randomized dormant EnKF is specifically designed for systems with limited uncertainty growth and extreme sampling error scenarios
- Results from toy, small-scale and large-scale realistic models suggest:
  (1) RD-EnKF is able to maintain sufficient ensemble spread after the update
  (2) RD-EnKF is robust even in highly-biased systems
  (3) RD-EnKF makes use of small ensemble sizes more efficiently than the EnKF
  (4) RD-EnKF requires less inflation (and localization) than the EnKF, given the inherent spread retention by the dormant members

# Conclusion

- A new method for tackling sampling errors is proposed
- The randomized dormant EnKF is specifically designed for systems with limited uncertainty growth and extreme sampling error scenarios

- Results from toy, small-scale and large-scale realistic models suggest:
  (1) RD-EnKF is able to maintain sufficient ensemble spread after the update
  (2) RD-EnKF is robust even in highly-biased systems
  (3) RD-EnKF makes use of small ensemble sizes more efficiently than the EnKF
  (4) RD-EnKF requires less inflation (and localization) than the EnKF, given the inherent spread retention by the dormant members

- Mixture likelihood for non-Gaussian observations
- Different ways to randomly (or not) select the dormant members
- Ways to hange the dormancy rate between cycles?