# Data Assimilation Framework For Earth System Models

**Andrew Stuart**

Shiwei Lan, Tapio Schneider, João Teixeira
California Institute of Technology

**ONR N00014-17-1-2079**
**Caltech-JPL President's and Director's Fund**
**Charles Trimble**

May $17^{th}$ 2018

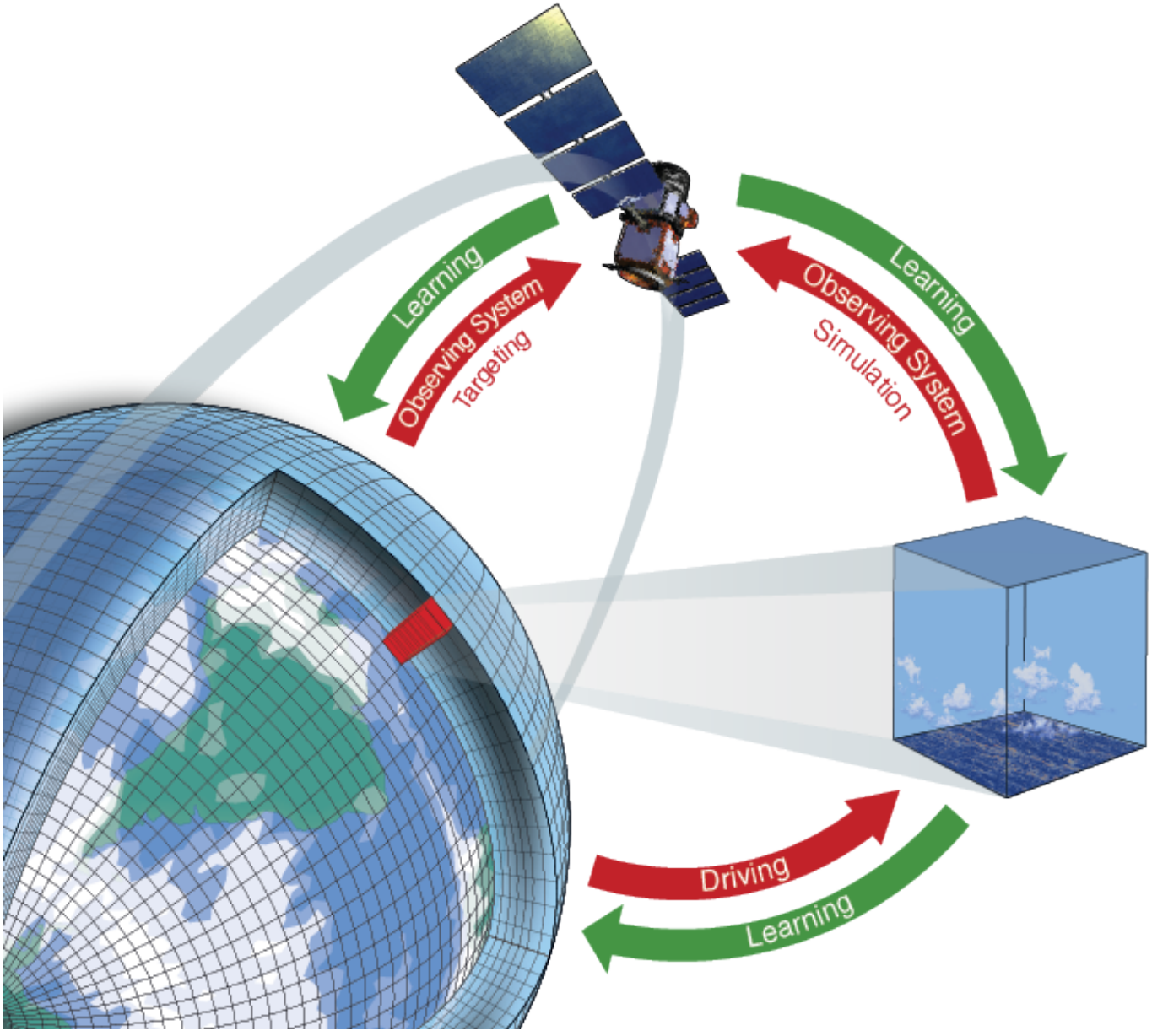



Figure: Satellite and HRS (High Resoution Simulation) Data

# Overview

Overarching Themes

Research Program

Conclusions

# Overarching Themes

# Motivation

- Major source of uncertainty in climate models is from sub-grid:
  - Atmosphere: clouds;
  - Ocean: small scale turbulence;
  - Biosphere, cryosphere ....
- Sub-grid model parameters are not directly measurable.
- Data: satellite (global), drifters (global) and simulation (local).
- Improving parameter estimation is crucial.
- Quantifying and reducing uncertainty in estimation is crucial.

# Key Drivers

- Physics-based models are needed for predictive capability.
- Differing data sources should be co-integrated.
- Forward model runs limited by computational cost.
- Machine learning tools for judicious surrogates.

# Key Ideas

- Time-averaged data (avoids initialization; smooths objective). ✓
- Ensemble Kalman inversion (no derivatives; few forward runs). ✓
- ML surrogates to enhance ensembles and perform Bayesian UQ. ✓
- Mini-batching of data to develop on-line filtering methods. ✗
- ML informed experimental design to locate HRS. ✗

✓: trialled for parameter learning/UQ for Lorenz 96.

✗: ideas to be developed.

# Research Program

# Test Problem

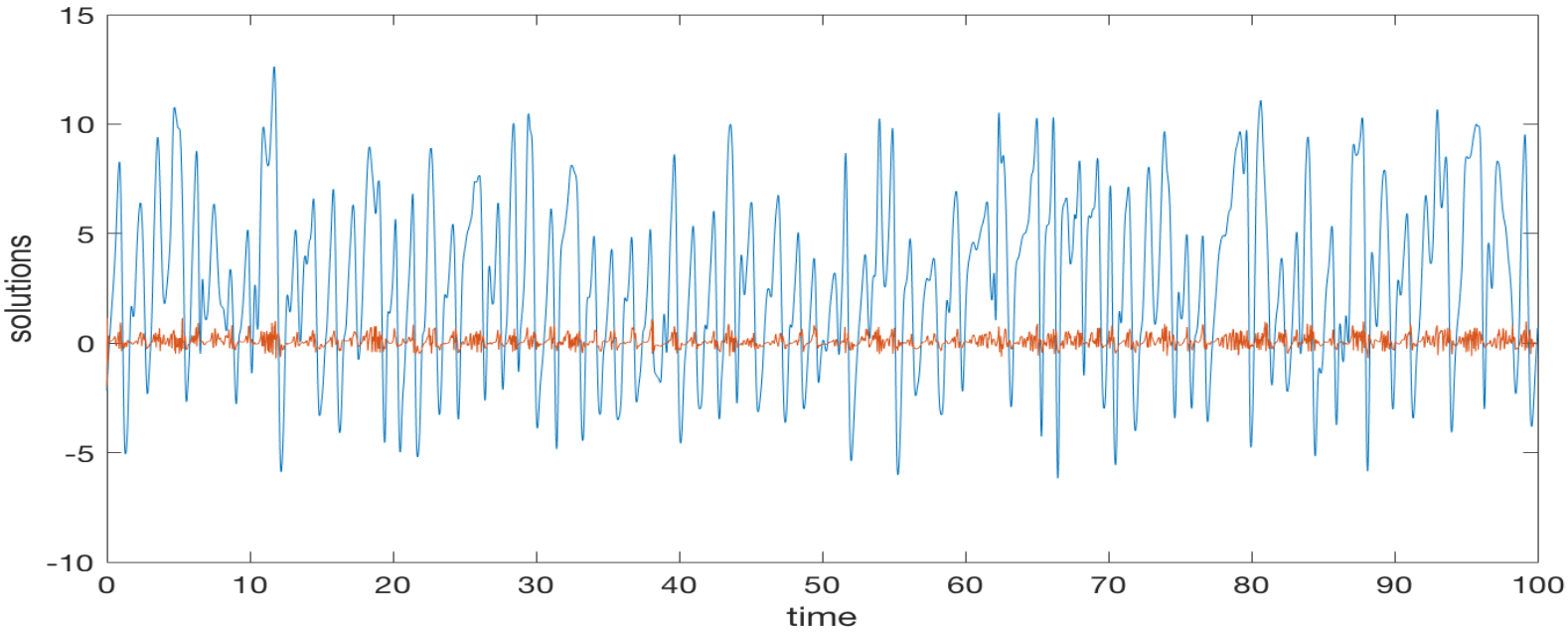



Figure: Multi-scale atmospheric dynamics model of Lorenz (1996).

$$\frac{dX_k}{dt} = -X_{k-1}(X_{k-2} - X_{k+1}) - X_k + F - hc\bar{Y}_k,$$

$$\frac{1}{c}\frac{dY_{j,k}}{dt} = -bY_{j+1,k}(Y_{j+2,k} - Y_{j-1,k}) - Y_{j,k} + \frac{h}{J}X_k.$$

# Parameter Estimation

Find parameters $\theta = (F, h, c, b)$ from data $y$ linked putatively by

$$\boxed{y = \mathcal{G}(\theta) + \text{noise}.}$$

Use time-averages to smoothen the objective:

$$\boxed{\mathcal{G}(\theta) = \frac{1}{T}\int_0^T \phi\big(z(t;\theta)\big)\,dt, \quad y = \frac{1}{T}\int_0^T \phi\big(z^\dagger(t)\big)\,dt.}$$

Computing $z(t;\theta)$ involves running expensive forward model.
$z^\dagger(t)$ is unprocessed high frequency data.
$y$ is processed data that we train $\theta$ on.

# Ensemble Kalman Inversion (EKI)

Evenesen (1994); Anderson (2001); Oliver, Reynolds and Liu (2008).

**Algorithm 1** Continuous Time EKI

1. *Sample parameters* $\{\theta^{(j)}(0)\}_{j=1}^{M}$ *from prior.*
2. *Solve coupled ODEs*

$$\dot{\theta}^{(j)} = -\sum_{\ell=1}^{M} d^{(j,\ell)}\theta^{(\ell)},$$

$$d^{(j,\ell)} = \frac{1}{M}\sum_{k=1}^{M}\left\langle \mathcal{G}(\theta^{(j)}) - y, \mathcal{G}(\theta^{(\ell)}) - \mathcal{G}(\theta^{(k)})\right\rangle$$

Derivative free; drives to consensus; drives to data.

# EKI: Parameter Estimation

Schnieder, Lan, Stuart, Teixeira
(Geo. Res. Lett. 2017, arxiv.1709.00037).

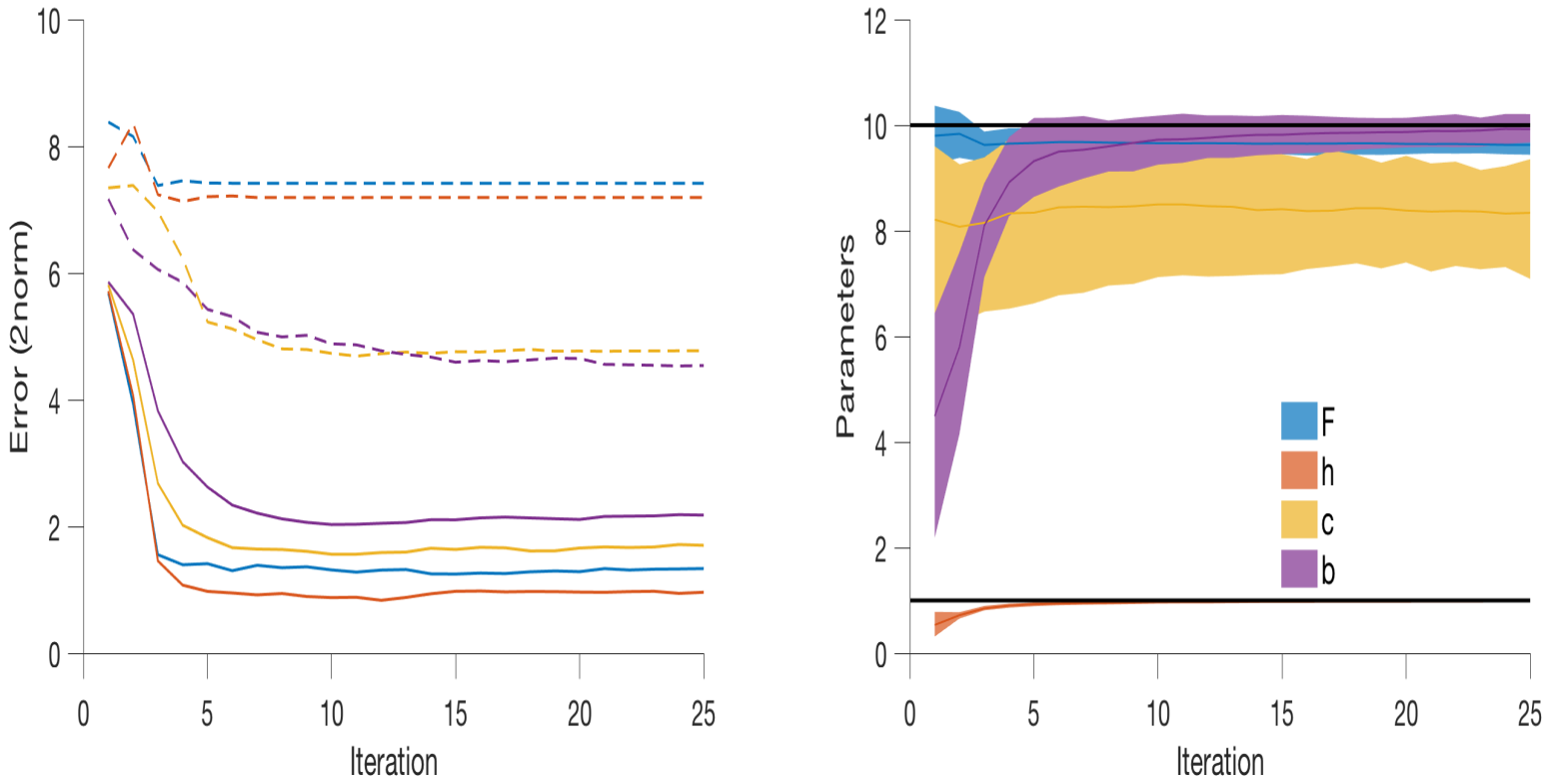



Figure: Left: parameter estimates $M = 10$ (dash) and $M = 100$ (solid).
Right: ensemble spread for $25 - 75\%$ quartile when $M = 100$.

# EKI: Errror Versus Iteration Number

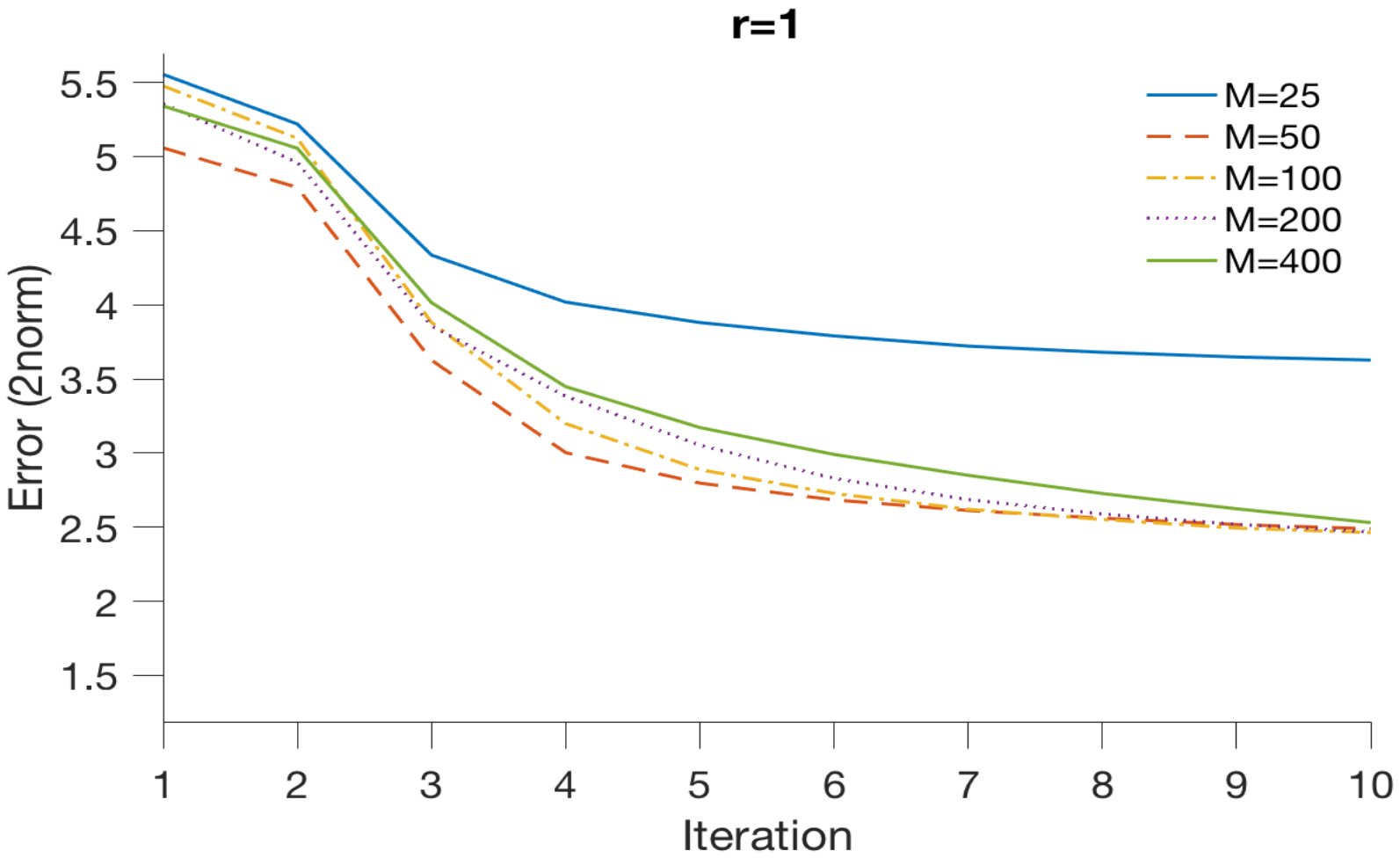



Figure: $\|\theta^* - \hat{\theta}_{\mathrm{EnKF}}\|_2$, versus iteration number.

# EKI: Errror Versus Ensemble Size

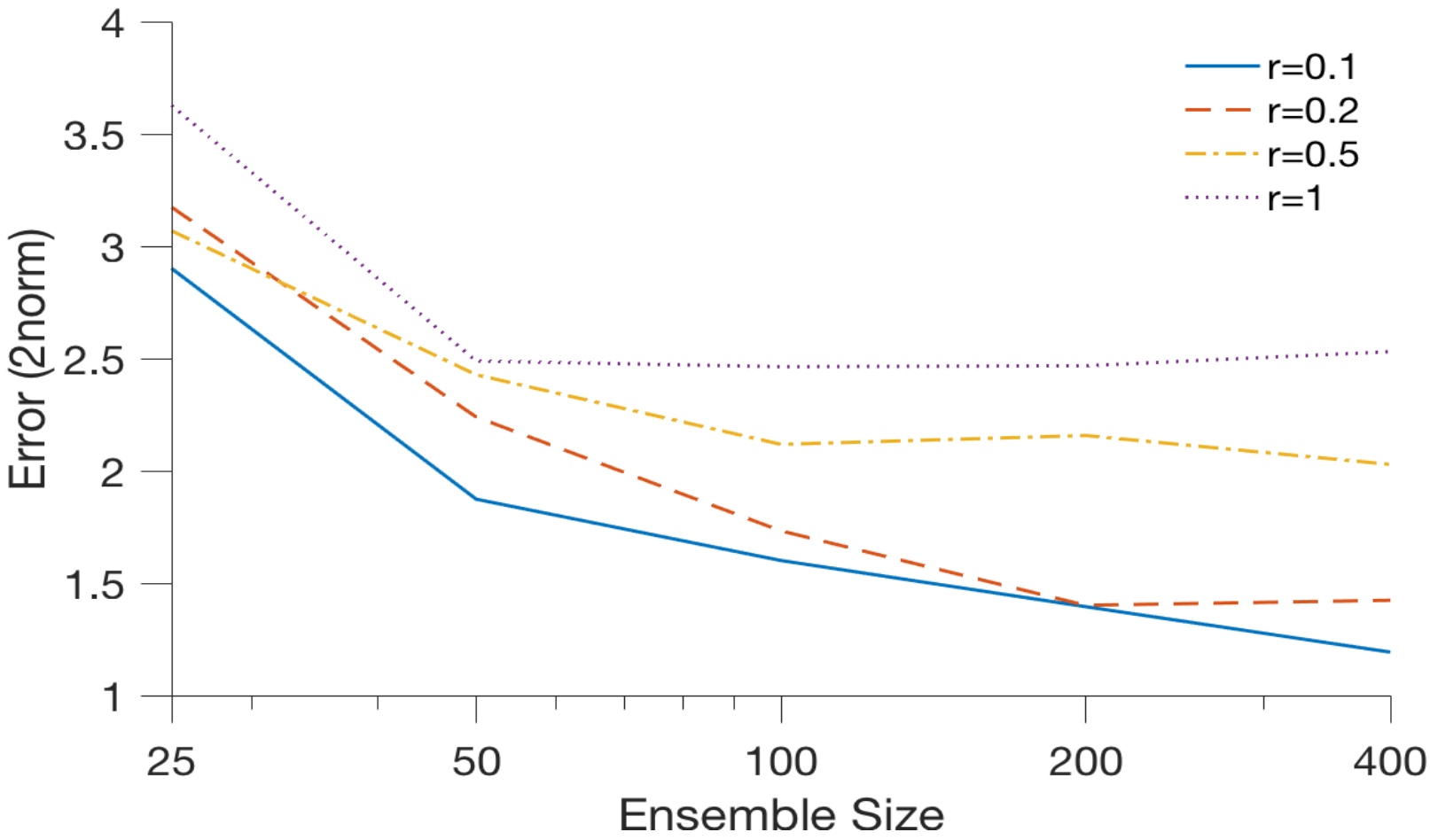



Figure: $\|\theta^* - \hat{\theta}_{\mathrm{EnKF}}\|_2$, versus ensemble size $M$. $10^{th}$ iteration.

# EKI: Relative Errror in STD (c/w MCMC)

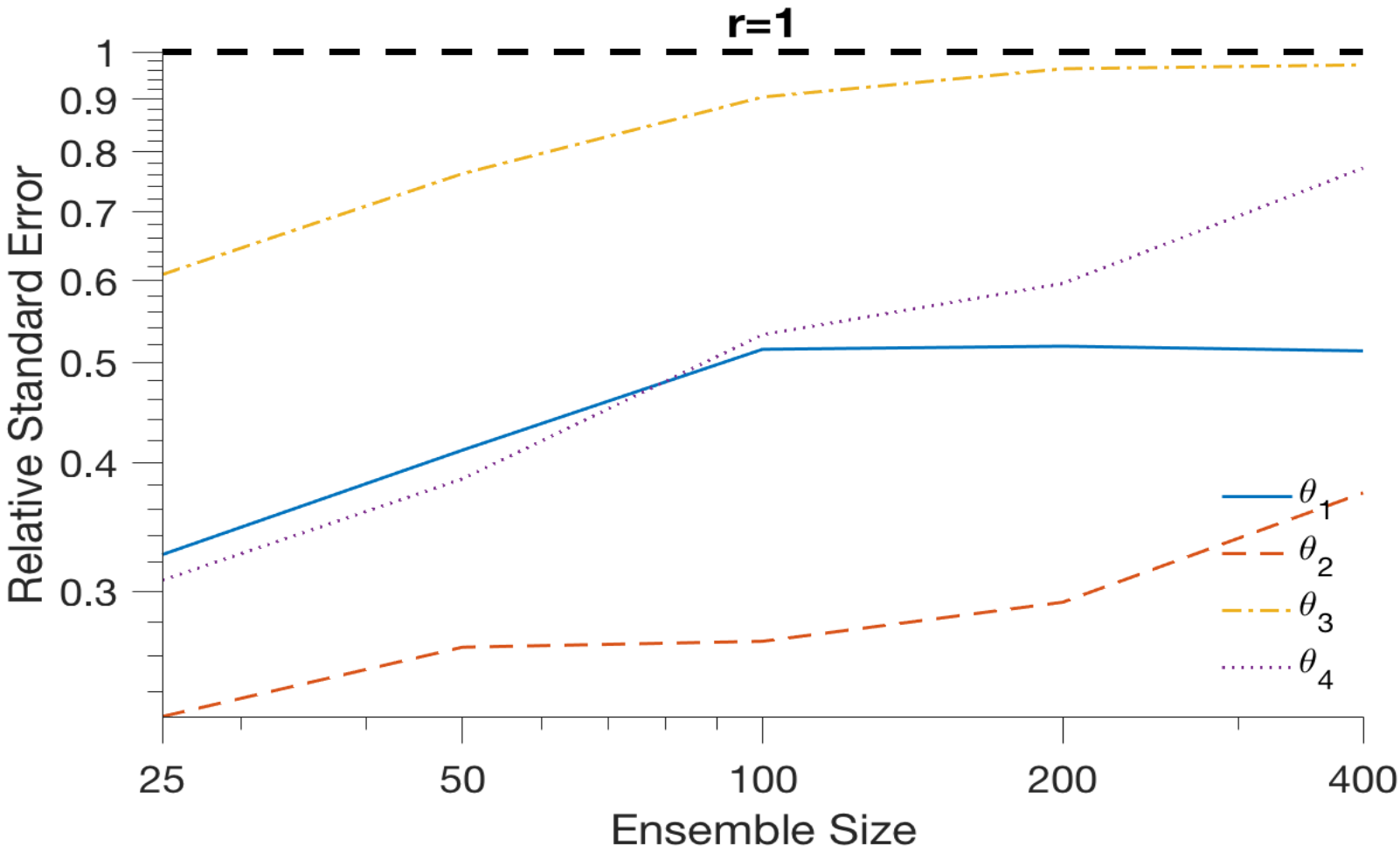



Figure: Relative error in standard deviation versus ensemble size $M$.

# Uncertainty Quantification

Issue:

- EKI for optimization uses $\mathcal{O}(10^2)$ model runs (acceptable).
- EKI underestimates uncertainty.
- MCMC for UQ uses $\mathcal{O}(10^6)$ model runs (unacceptable).

Solution:

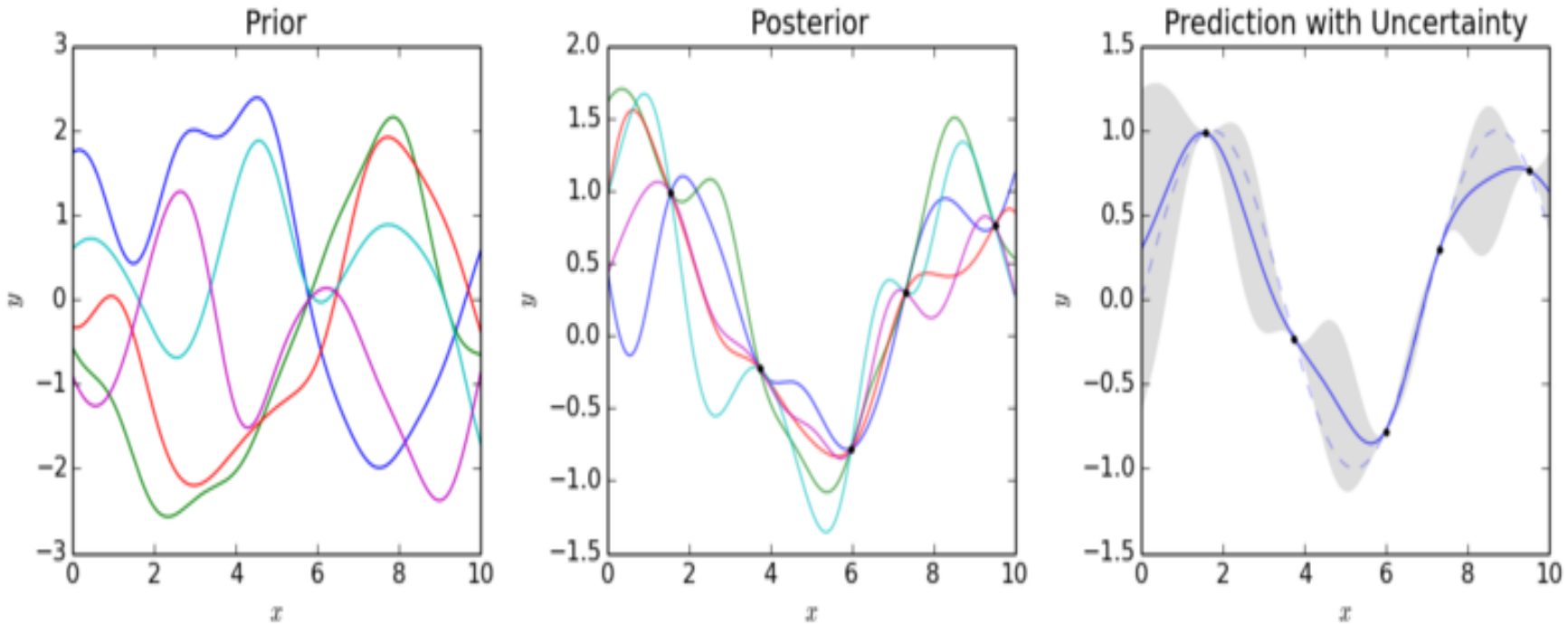
- Use GP (DNN?) to approximate model; train on EKI.
- Key: EKI spread to cover support of posterior.
- Theoretically justified. Stuart and Teckentrup (Math. Comp. 2018, `arxiv.1603.02004`)

# Gaussian Process Regression

**Prior:**

$$y(x_n) = \mathcal{G}(x_n) + \epsilon_n, \quad \epsilon_n \overset{iid}{\sim} \mathcal{N}(0, \sigma_y^2)$$
$$\mathcal{G}(x) \sim \mathrm{GP}(m(x), k(x, x'))$$
$$m(x) = 0, \quad k(x, x') = \sigma^2 \exp\{-\|x - x'\|^2/(2\ell^2)\}$$

**Posterior:** Condition on model runs from EKI.



# UQ: Results

Schnieder, Lan, Stuart. (In preparation, 2018).

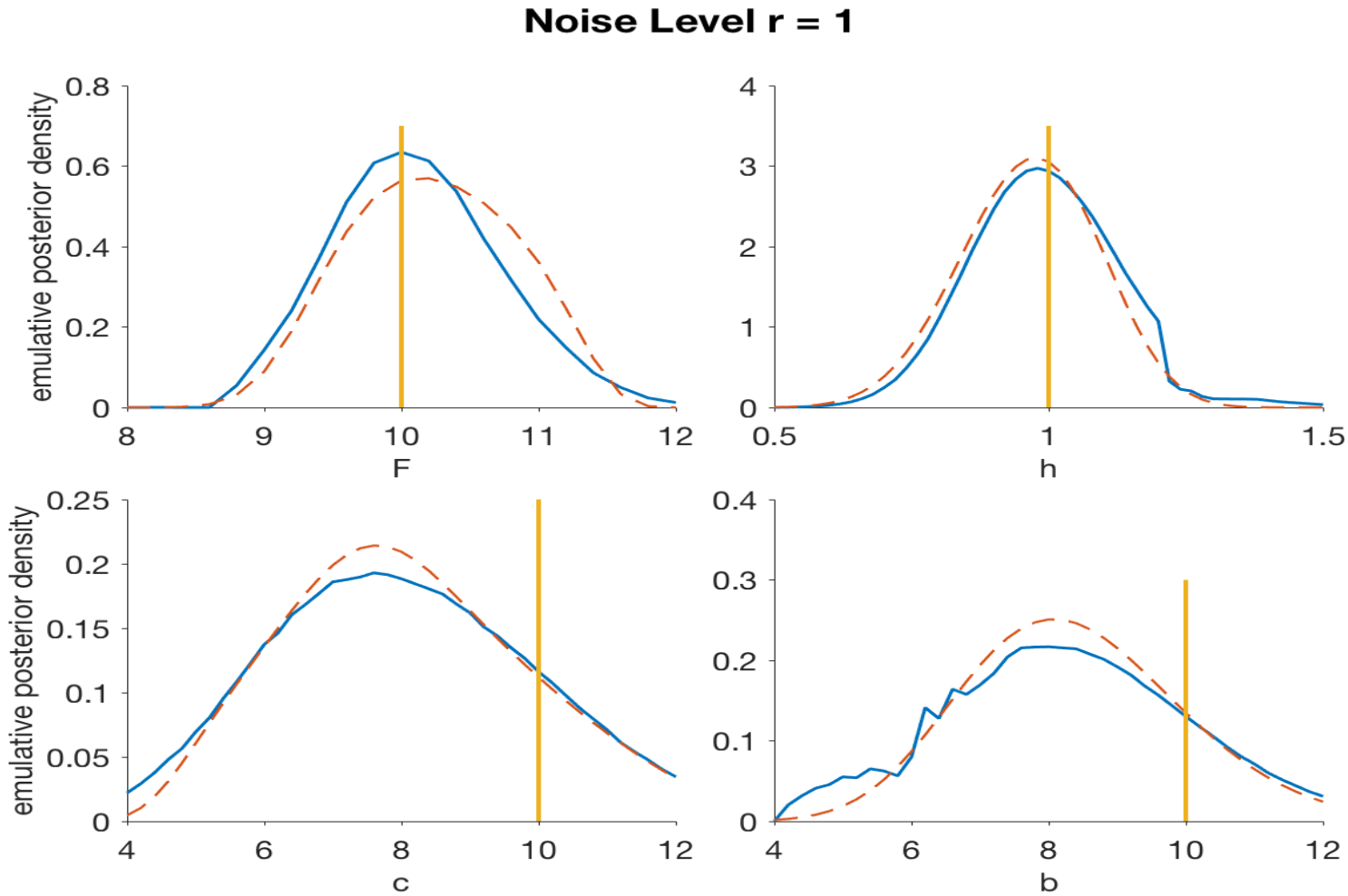



Figure: True posterior marginals (solid); emulated posterior marginals (dashed).

# Emulative MCMC: Relative Errror in STD (c/w MCMC)

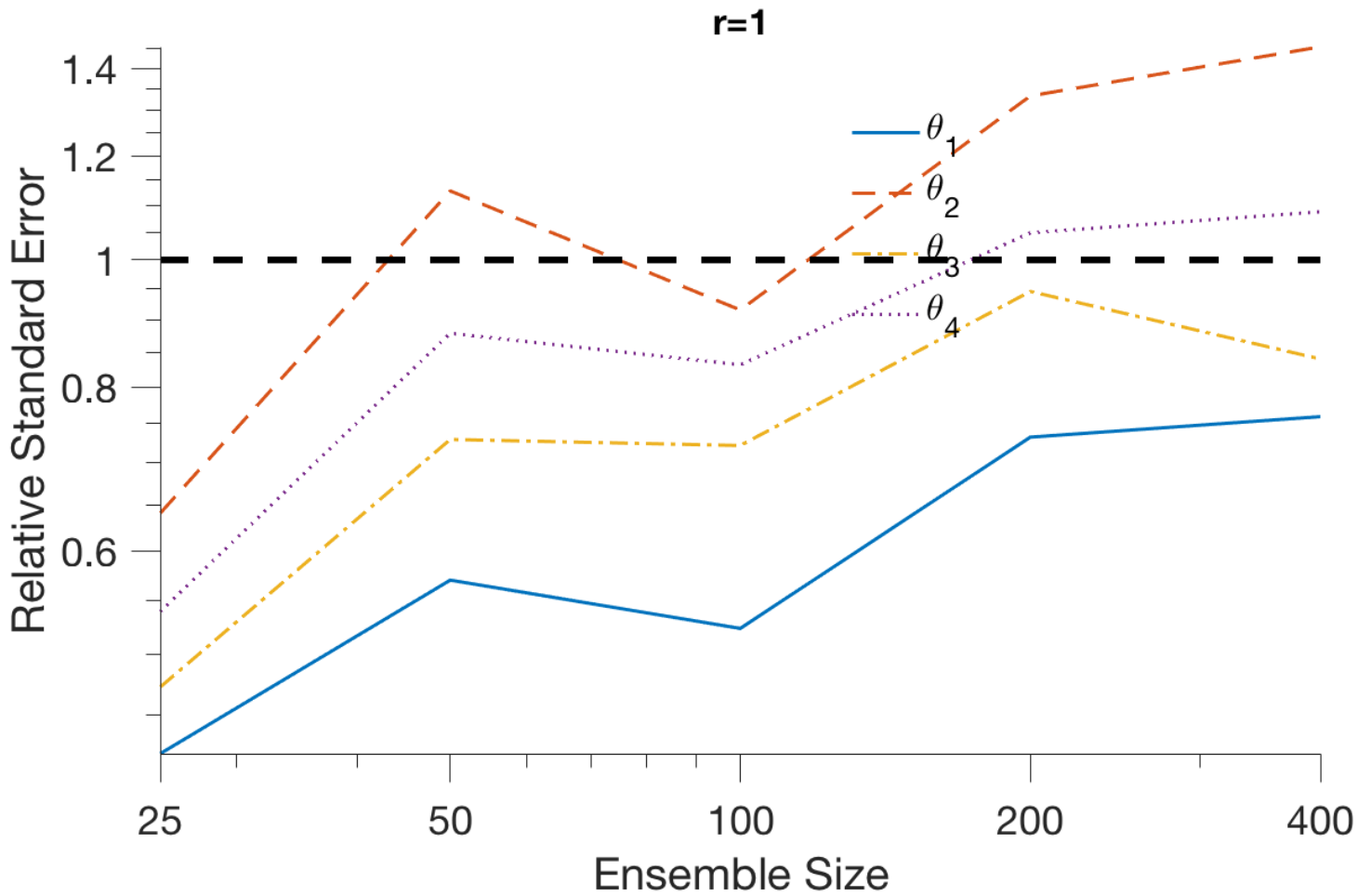



Figure: std (upper) and relative std (lower) of GP emulation as a function of $M$, with $n = M \times 4$ training points taken from the first 4 iterations respectively.

# On-line Filtering

Issue:

- EKI as described uses all time-ordered data at once.
- Breaking up time risks accuracy.
- Breaking up time gains efficiency.
- Smoothing versus filtering.

Proposed solution:

- Use mini-batching of data (as in training of DNNs).
- Employ in context of EKI not SGD.
- Experience with this approach for training CNNs on MNIST.

# Experimental Design

Issue:

- Cannot afford HRS simulations in every cloud column etc.
- Where to perform HRS is an experimental design problem.
- Sensitivities expensive to compute.

Proposed solution:

- Use ML to guide location of HRS.
- Use genetic algorithms (Metropolis) to guide location of HRS.

# Conclusions

# Conclusions

A robust, adaptive and data-informed climate model will be developed on a five year time-scale, guided by theory and small scale model-backed intuition. Key components are:

- Physics based predictive models.
- State of the art Data Assimilation.
- State of the art Inverse Problems.
- Judicious infusion of ML.