Data Assimilation in the L96 Model: A Route to Understanding Model Error

Feiyu Lu, Judith Berner, Mitch Bushuk

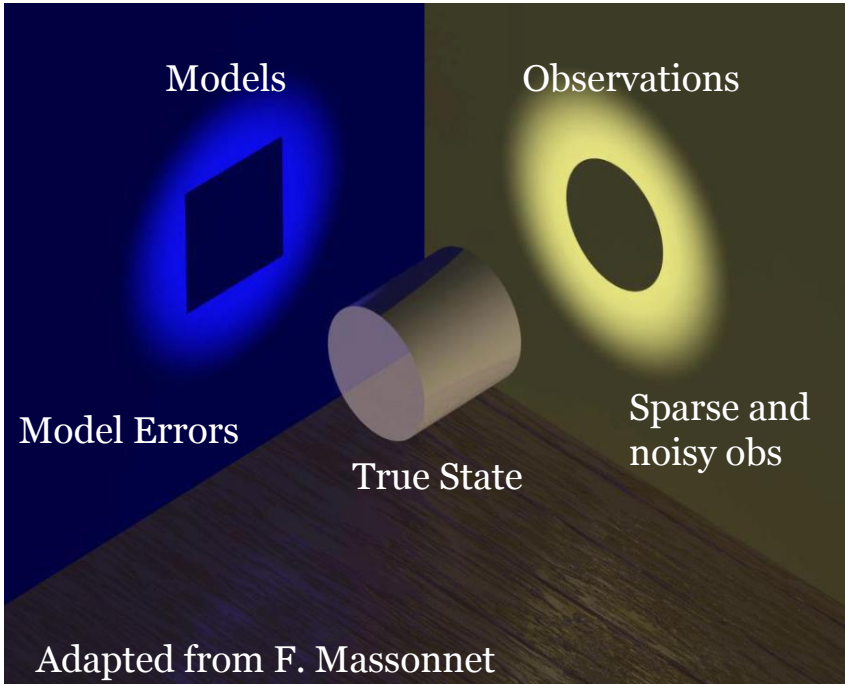
All Code and Jupyter Notebooks available here:
https://github.com/m2lines/L96_demo

Plan For Today's Demo:

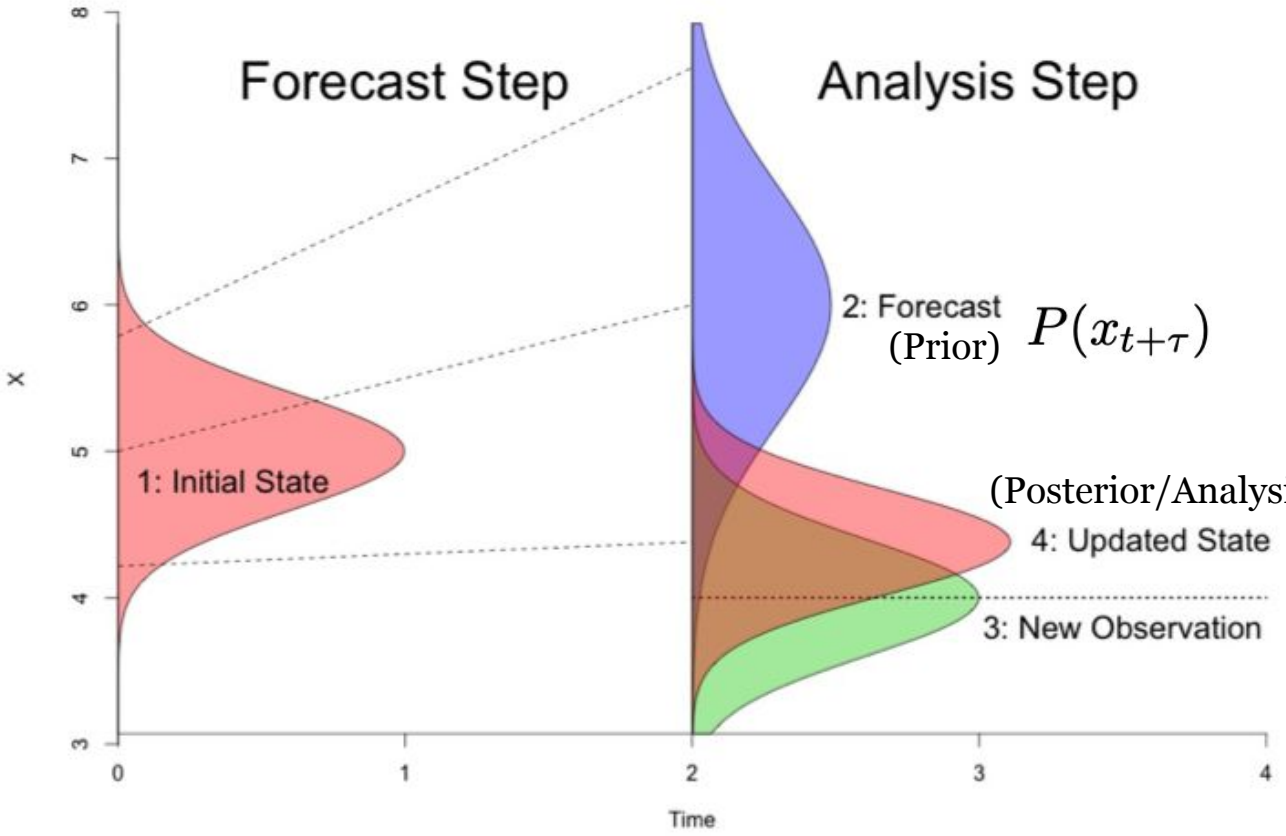
1. **Introduction to Data Assimilation (DA) and the L96 toy problem**
2. L96 DA code demo
3. Using DA to understand and improve L96 model error

What is DA? Why do we do it?

- Data assimilation is:
  - An approach for combining data (observations) with prior knowledge (from a physical model) to obtain an estimate of the true state of a system.
- Why do we do it?
  - Generate accurate initial conditions for weather and climate predictions
  - Produce reanalysis products (state estimates) of the atmosphere-ocean-sea ice-land system
  - M²LInES project: Understand and improve climate model errors
- Challenges
  - Observations are sparse, incomplete, and noisy representations of the true state
  - Models needed to “fill in” missing observational data, but have model errors

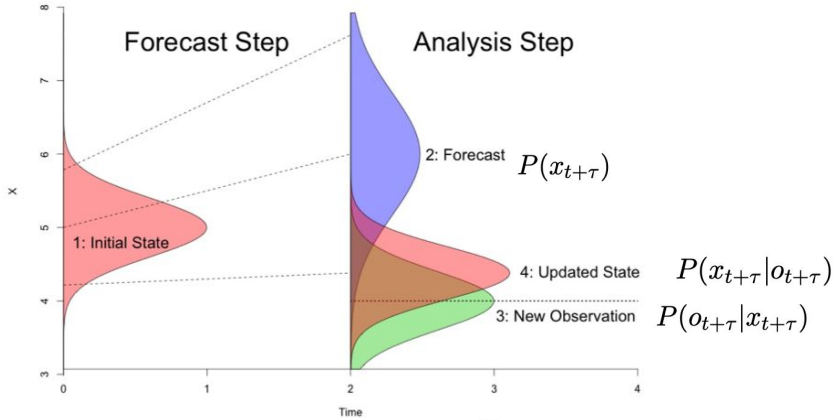


Sequential DA: Forecast and Analysis Steps



DA updates: Bayes' Rule

$$\underbrace{P(x_{t+\tau}|o_{t+\tau})}_{\text{Posterior}} = \frac{\overbrace{P(o_{t+\tau}|x_{t+\tau})}^{\text{Obs Likelihood}}\overbrace{P(x_{t+\tau})}^{\text{Prior}}}{\underbrace{P(o_{t+\tau})}_{\text{Marginal (normalization)}}}$$



$P(x_{t+\tau}|o_{t+\tau}) : \mu_p, \sigma_p^2$ Posterior

$P(x_{t+\tau}) : \mu_f, \sigma_f^2$ Prior

$P(o_{t+\tau}|x_{t+\tau}) : \mu_o, \sigma_o^2$ Obs Likelihood

In the case of Gaussian prior and observational distributions, posterior can written as:

$$\mu_p = \frac{\sigma_f^2}{\sigma_f^2 + \sigma_o^2}\mu_o + \frac{\sigma_o^2}{\sigma_f^2 + \sigma_o^2}\mu_f \qquad \sigma_p^2 = \frac{1}{\frac{1}{\sigma_f^2} + \frac{1}{\sigma_o^2}}$$

DA updates: Increments and Gain

$$\mu_p = \frac{\sigma_f^2}{\sigma_f^2 + \sigma_o^2}\mu_o + \frac{\sigma_o^2}{\sigma_f^2 + \sigma_o^2}\mu_f \qquad \sigma_p^2 = \frac{1}{\frac{1}{\sigma_f^2} + \frac{1}{\sigma_o^2}}$$

Rearranging, we can write:

$$\mu_p = \mu_f + \underbrace{K(\mu_o - \mu_f)}_{\text{increment}} \qquad \sigma_p^2 = (1-K)\sigma_f^2$$

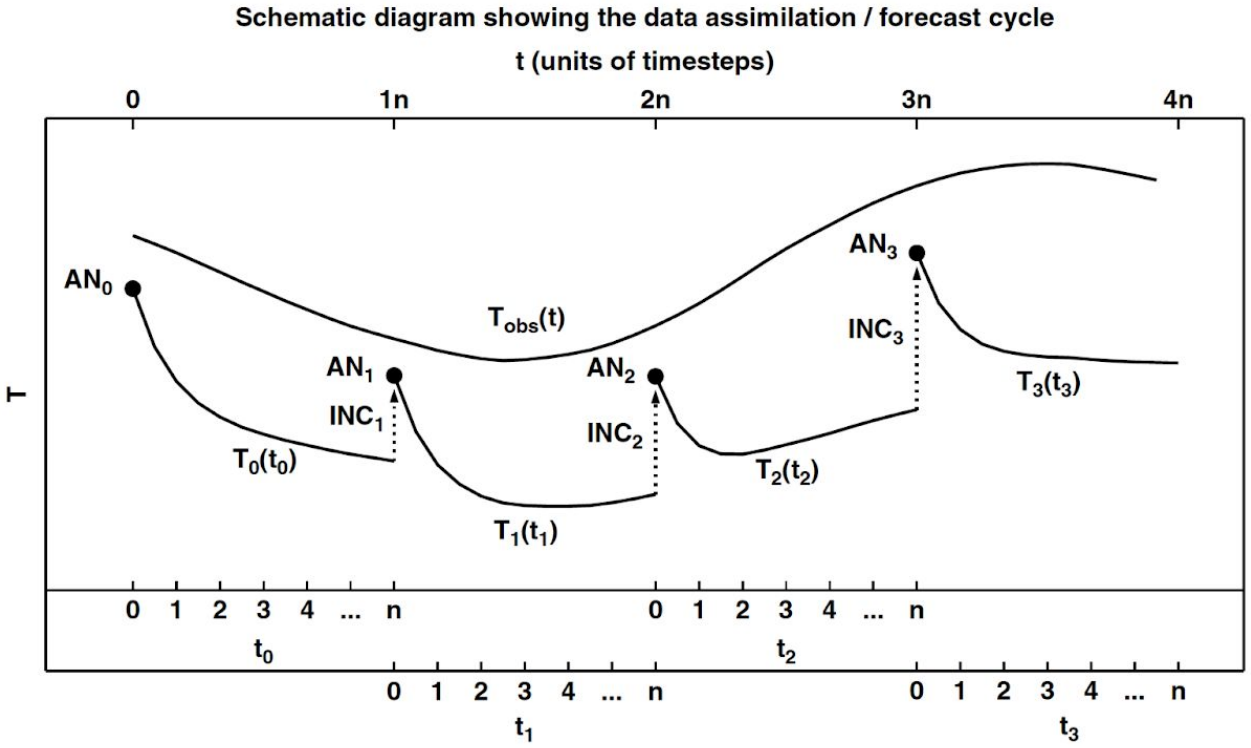
where,

$$K = \frac{\sigma_f^2}{\sigma_f^2 + \sigma_o^2}$$

is referred to as the “gain” or “Kalman gain” ($0<K<1$).

These 1-D expressions can be generalized to multi-dimensional systems (variances become covariance matrices). This allows for unobserved state variables to be updated via observations of another state variable, which they covary with.

Data Assimilation Increments

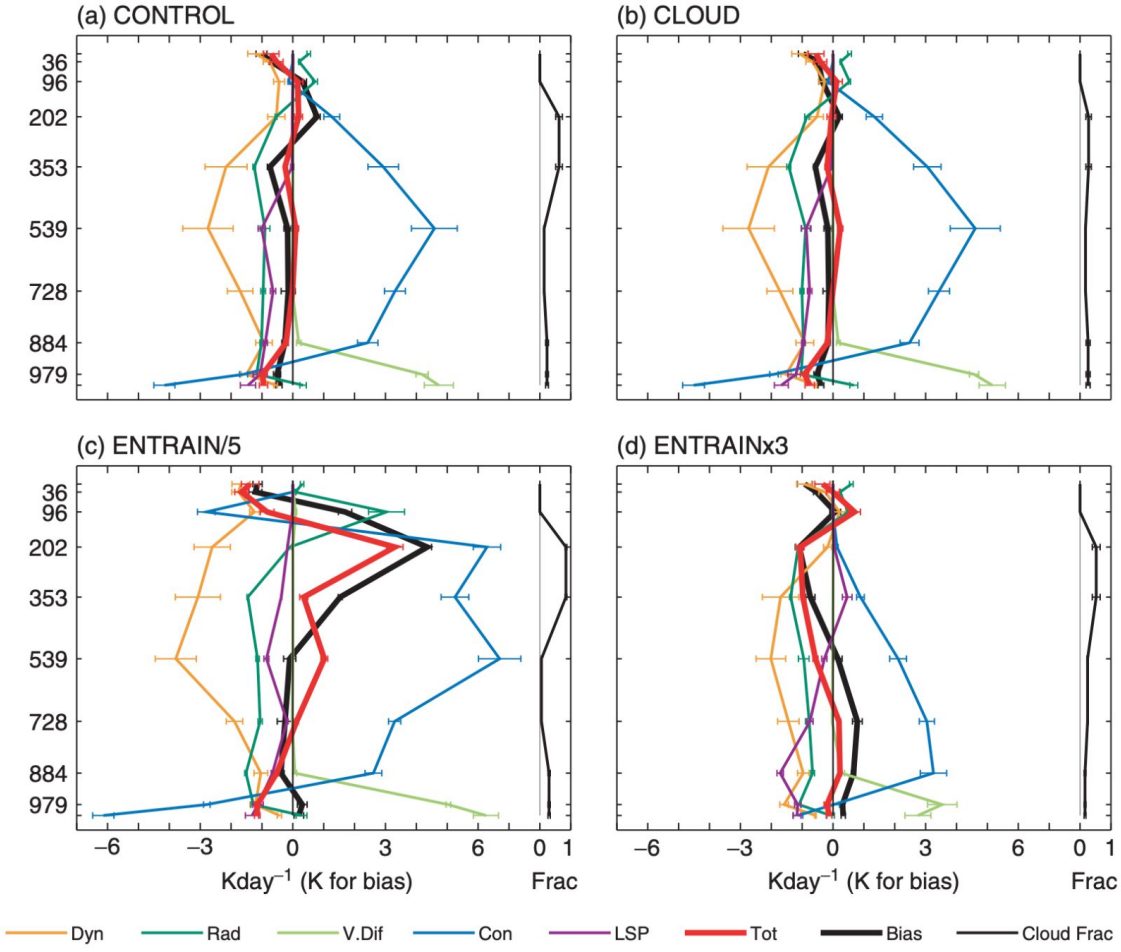


The analysis increment is consistently positive, which is evidence for model being systematically too cold.

Figure 1. Schematic diagram showing the data assimilation and forecast integration aspects of numerical weather prediction. $T_{\mathrm{obs}}(t)$ represents an observed time series (e.g. of temperature at some specified location). For each i, $T_i(t_i)$ represents the model forecast initiated from analysis AN_i. For the purposes of explaining our methodology, the role of systematic forecast error (in this case a cooling) has been emphasized over random error. See the main text for further explanation.

Rodwell and Palmer 2007

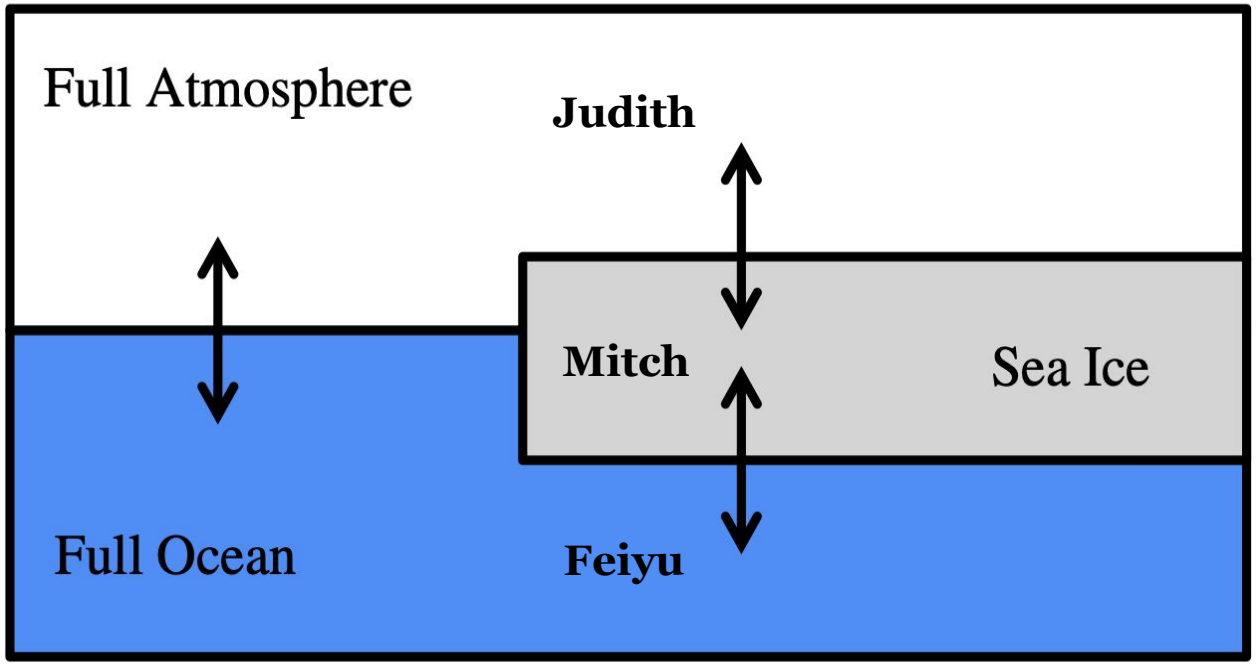
Using DA to Understand and Quantify Model Error



Rodwell and Palmer 2007

- Used DA runs to examine model error
- Total temperature tendencies (thick red) from DA run mirror model bias (thick black)
- Decomposition of tendencies into process-based contributions provides physical insights on model error
- Bottom: changing parameter “entrainment rate” leads to large systematic DA increments

M²LInES: Using DA to Understand and Quantify Model Error



- Use DA systems to improve model errors in atmosphere (Judith), ocean (Feiyu), and sea ice (Mitch) components
- Connect DA increments with model error and physical processes
- Develop ML models to learn state-dependent increments
- Develop and implement climate model parameterizations based on ML models.

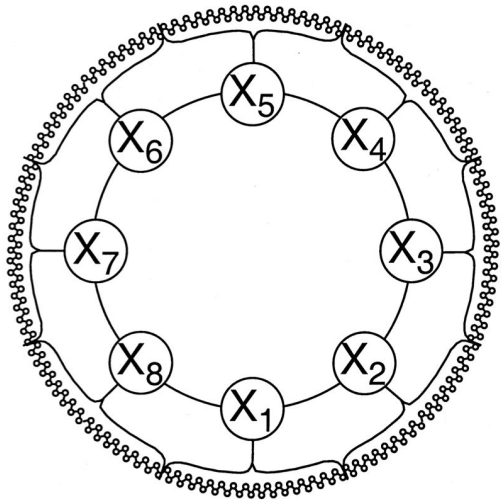
L96 DA Toy Problem Setup

The "Truth": 2-scale L96 Model

$$\frac{d}{dt}X_k = -X_{k-1}(X_{k-2} - X_{k+1}) - X_k + F - \left(\frac{hc}{b}\right)\sum_{j=0}^{J-1} Y_{j,k}$$

$$\frac{d}{dt}Y_j, k = -cbY_{j+1,k}(Y_{j+2,k} - Y_{j-1,k}) - Y_{j,k} + \frac{hc}{b}X_k$$

$$\text{BC} : X_{k+K} = X_k; \quad Y_{j,k+K} = Y_{j,k}; \quad Y_{j+J,k} = Y_{j,k+1}$$



Wilks 2005

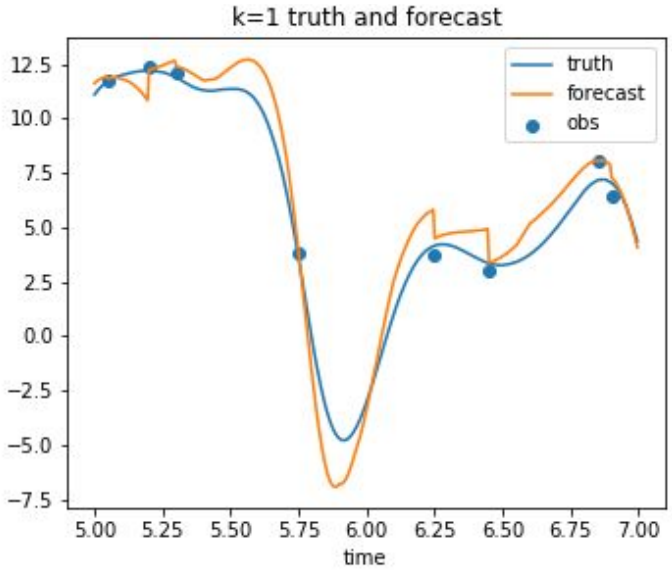
The "GCM": 1-scale L96 Model w/ Parameterization

$$\frac{d}{dt}X_k = -X_{k-1}(X_{k-2} - X_{k+1}) - X_k + F_{GCM} + \underbrace{P(X_k)}$$

Parameterization based on coarse resolution state variables

L96 DA Toy Problem Setup

1. Generate a “truth” signal from 2-scale L96 model
2. Specify a biased “GCM”model for performing DA
   a. Choose P (to start, use GCM with P=0).
   b. Choose F_GCM (to start, use F_GCM=F).
3. Collect sparse and noisy observations, sampled from “truth”
   a. Specify observation density in space, frequency in time
   b. Specify observational error
4. Perform DA using a method of choice
   a. Choose DA method: EnKF, 3DVAR, Hybrid EnKF
   b. Specify DA setup: ensemble members, localization, inflation
5. Assess DA performance
   a. RMSE of DA experiment relative to truth
6. Analyze DA increments to assess GCM model error



L96 DA Toy Problem Setup: Sources of Forecast Error

1. **Errors in Forcing**
   a. Constant offset: $F_{GCM} = F + c$
   b. Spatial structure in true forcing (e.g. subgrid orography): $F := F(\mathbf{X}, t)$
   c. Temporal structure in true forcing (e.g. seasonal cycle, trend; can be used for "out of sample" tests): $F := F(\mathbf{X}, t)$

2. **Errors in Physics**
   a. Unresolved and unrepresented processes: $P = 0$
   b. Errors in parameterization: $P(X_k) \neq -\left(\frac{hc}{b}\right) \sum_{j=0}^{J-1} Y_{j,k}$

3. **Numerical Errors**

4. **Initial Condition Errors**

Overall goal: Use DA increments to diagnose and improve model error

Plan For Today's Demo:

1. Introduction to Data Assimilation (DA) and the L96 toy problem
2. **L96 DA code demo**
3. Using DA to understand and improve L96 model error

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

In the context of DA:

- A: estimate of the true state
  - **analysis/posterior** on the left
  - **forecast/prior** on the right
- B: observation

- **P(B|A)**: probability distribution of observation given prior model estimate
  - Forward operator (model space to observation space)
  - Observation error
- **P(A)**: probability distribution of prior estimate of the true state (**forecast**)
  - Numerical/analytical calculation (KF for linear, Extended KF for nonlinear)
  - Climatological estimate (3DVar)
  - Ensemble estimate (EnKF)
  - Tangent Linear Model (4DVar) and its adjoint

Plan For Today's Demo:

1. Introduction to Data Assimilation (DA) and the L96 toy problem
2. L96 DA code demo
3. **Using DA to understand and improve L96 model error**

Detecting model error through DA

- The parameterization problem arising from the need to truncate the model equations
  - Structural model error
  - Case1: Bare truncation
  - Case2: With state-dependent parameterization
  -
- Parameter error
  - Parameter error is often indistinguishable from structural model error in complex models (e.g. entrainment example)
  - The forcing F of the two-timescale model is different from the L96-CGM - constant bias.

- Error in forcing
  - SST gradient in tropical Pacific is incorrect, leading to error in Walker circulation
  - Projection of missing subgrid scale orography on resolved scales
  - Add spatial variation to F in two-timescale model, but not L96-CGM
  - Formally a parameter error (if F is considered a parameter)

Detecting model error through DA

- Temporally varying error
  - Seasonally varying process
  - Add temporal variation to F in two-timescale model, but not L96-CGM

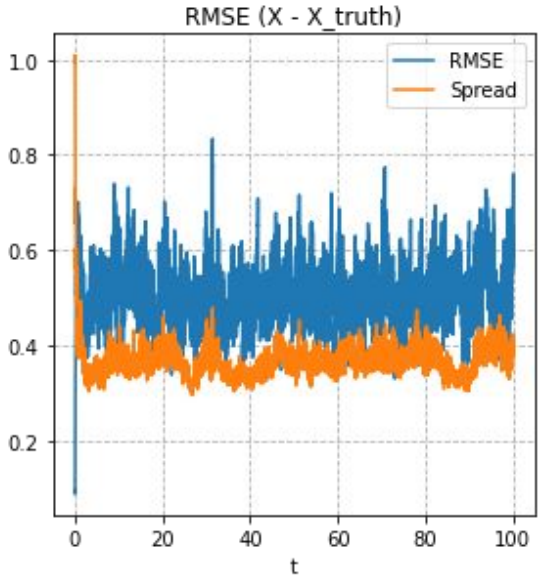
- “Flow-dependent” bias
  - Plant-Craig scheme does well for weakly forced convection, but not strongly forced one
  - Marine nocturnal PBL height too shallow

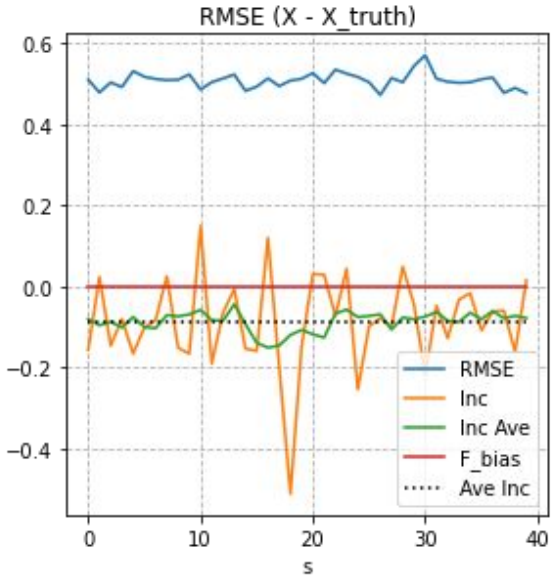
- Parameter estimation using DA
  - Augment the state vector with parameters, then use the covariance between state and parameters to come up with a parameter estimate

The fundamental parameterization problem

- The L96-CGM is a truncated version of the true two-timescale model
- Structural model error
- Case1: Bare truncation, the small scale variables are not represented at all

$$\frac{d}{dt}X_k = -X_{k-1}(X_{k-2} - X_{k+1}) - X_k + F_{GCM} + \cancel{P(X_k)}$$





➔ Magnitude of X is systematically to high => biased model

➔ Average increment is negative

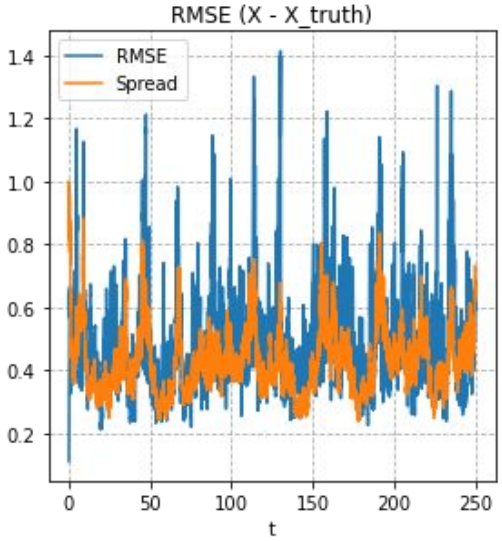
➔ DA increment needs to be linked to physical processes

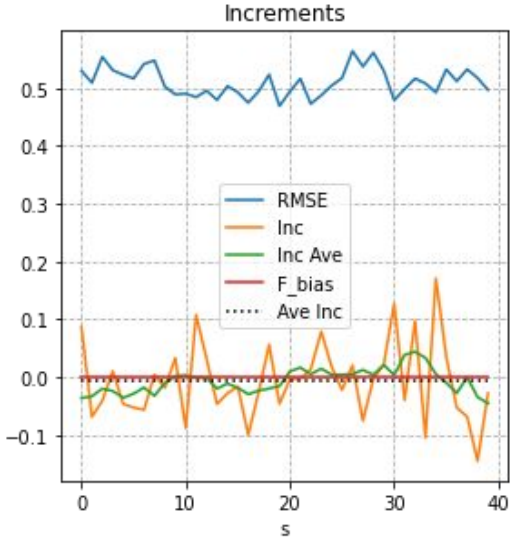
➔ “physical tendencies” can help to point at erroneous processes

The fundamental parameterization problem

- The L96-CGM is a truncated version of the true two-timescale model
- Structural model error
- Case2: State-dependent parameterization (polynomial fit)

$$\frac{d}{dt}X_k = -X_{k-1}(X_{k-2} - X_{k+1}) - X_k + F_{GCM} + P(X_k)$$





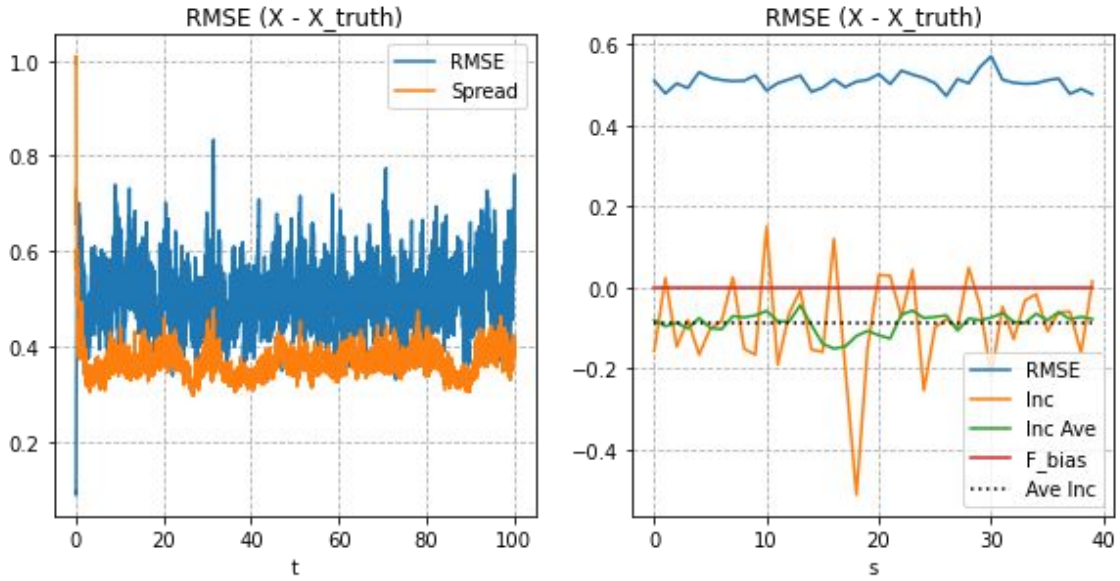
➔ Introducing parameterization removes model bias

➔ Average increment is close to zero

➔ Parameterization is so good, that mean DA-increment does not detect model bias

Parameter error

- Parameter error is often indistinguishable from structural model error in complex models (e.g. entrainment example)
- The forcing F of the two-timescale model is different from the L96-CGM



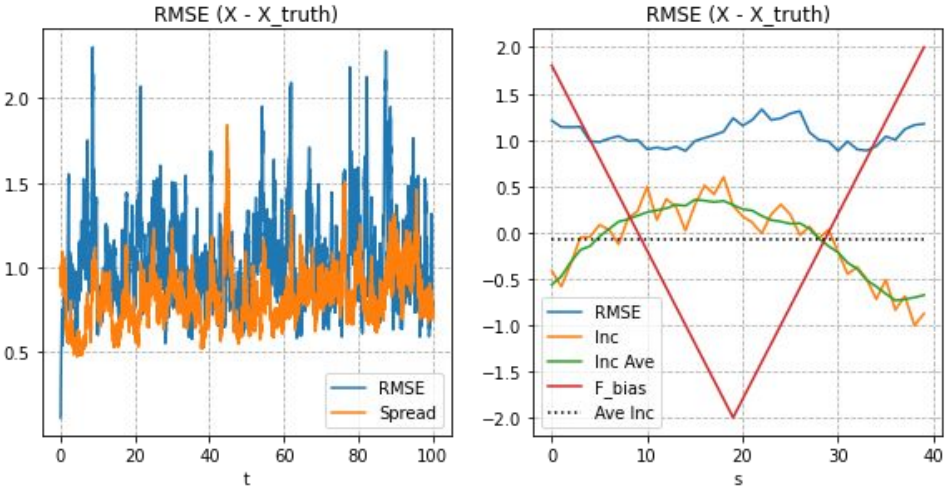
➔ Magnitude of X is systematically to high => biased model

➔ Average increment is negative

➔ Degeneracy

Spatial error in forcing

- SST gradient in tropical Pacific is incorrect, leading to error in Walker circulation
- Projection of missing subgrid scale orography on resolved scales
- Add spatial variation to F in two-timescale model, but not L96-CGM



➔ Analysis increment depends picks up spatial variation of forcing

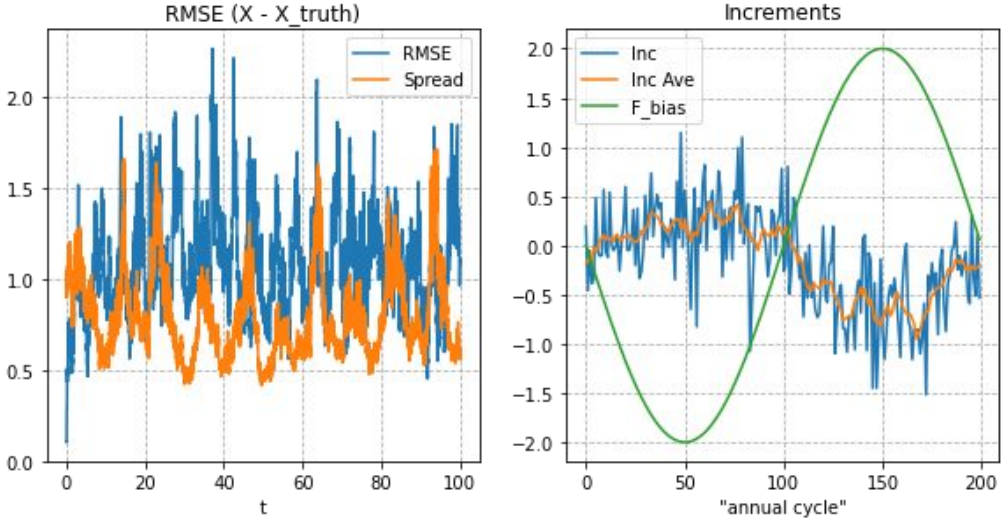
➔ Positive where forcing negative and vice versa

➔ Increment does not capture full magnitude of forcing variation

➔ Latter can be influenced by DA settings

Temporally varying error (e.g. seasonally varying parameter)

- Seasonally varying process
- Add temporal variation to F in two-timescale model, but not L96-CGM



➔ Analysis increment depends picks up temporal variation of forcing

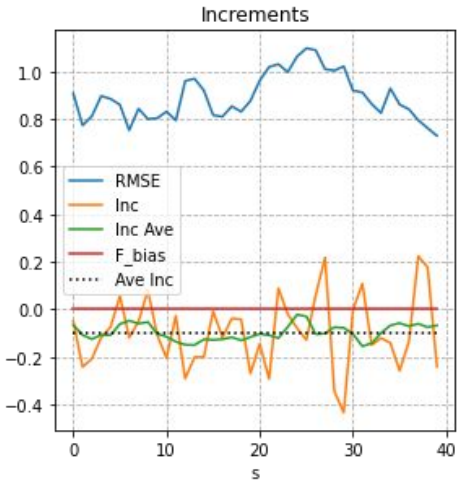
➔ Positive where forcing negative and vice versa

➔ Increment does not capture full magnitude of forcing variation

➔ Latter can be influenced by DA settings

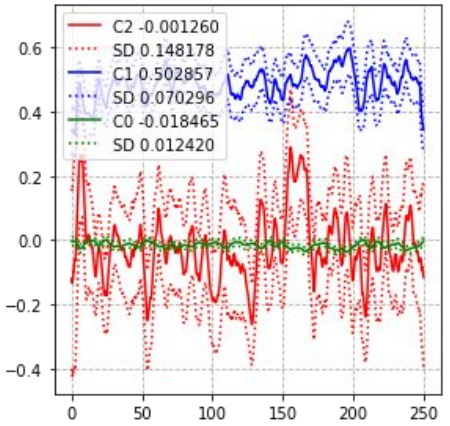
Parameter estimation - model error representation

Bare truncation

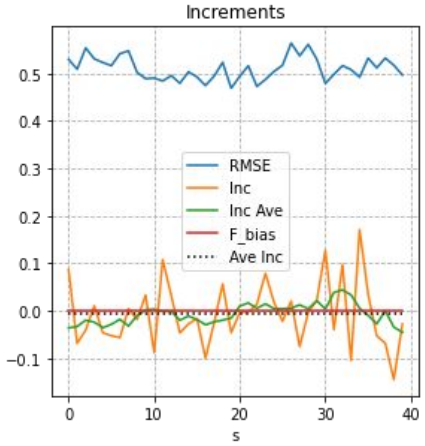


- ➔ Augment the state vector with parameters, then use the covariance between state and parameters to come up with a parameter estimate
- ➔ Posterior shows estimated parameter value and uncertainty
- ➔ Some parameter estimates converge converge more than others
- ➔ RMSE error reduced from 0.9 to 0.5 in simulations with learned parameterization
- ➔ Also reflected in reduced DA increment

Parameter estimation

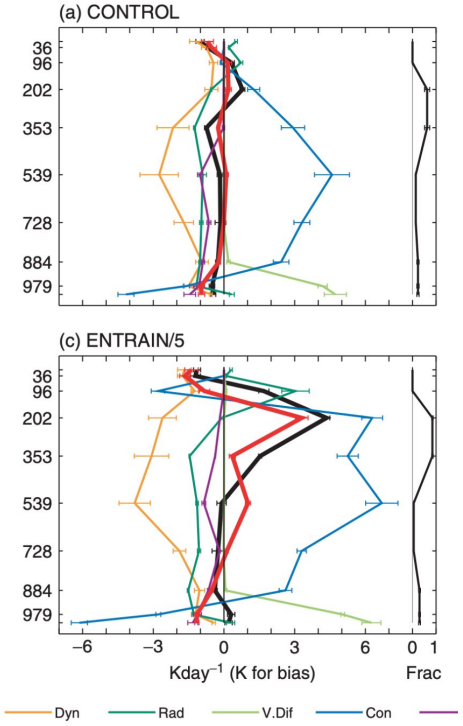


With learned parameters in polynomial parameterization



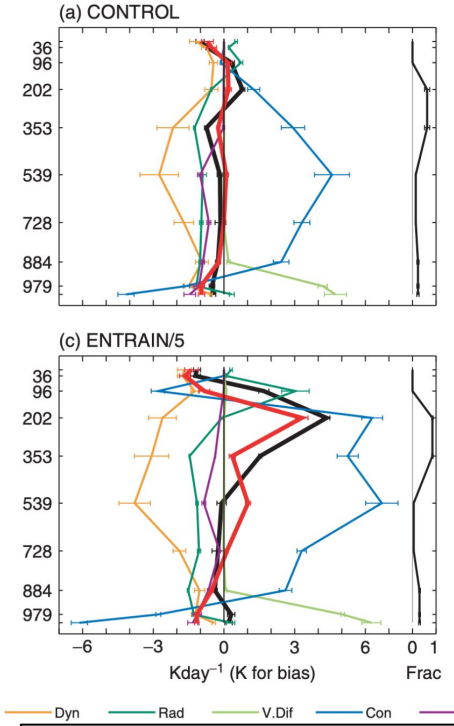
Is DA the holy grail for solving the model error problem?

- DA-increments are useful to detect model errors, especially systematic biases
- Degeneracy (different model errors can lead to the same signal in the DA-increment)
- Still needs domain expertise (e.g. link to physical tendencies) to understand physical sources
- Does not solve problem of complementing model errors



Is DA the holy grail for solving the model error problem?

- DA-increments are useful to detect model errors, especially systematic biases
- Degeneracy (different model errors can lead to the same signal in the DA-increment)
- Still needs domain expertise (e.g. link to physical tendencies) to understand physical sources
- Does not solve problem of complementing model errors



Rodwell and Palmer, 2008, see also Klinker and Sardeshmukh, 1997