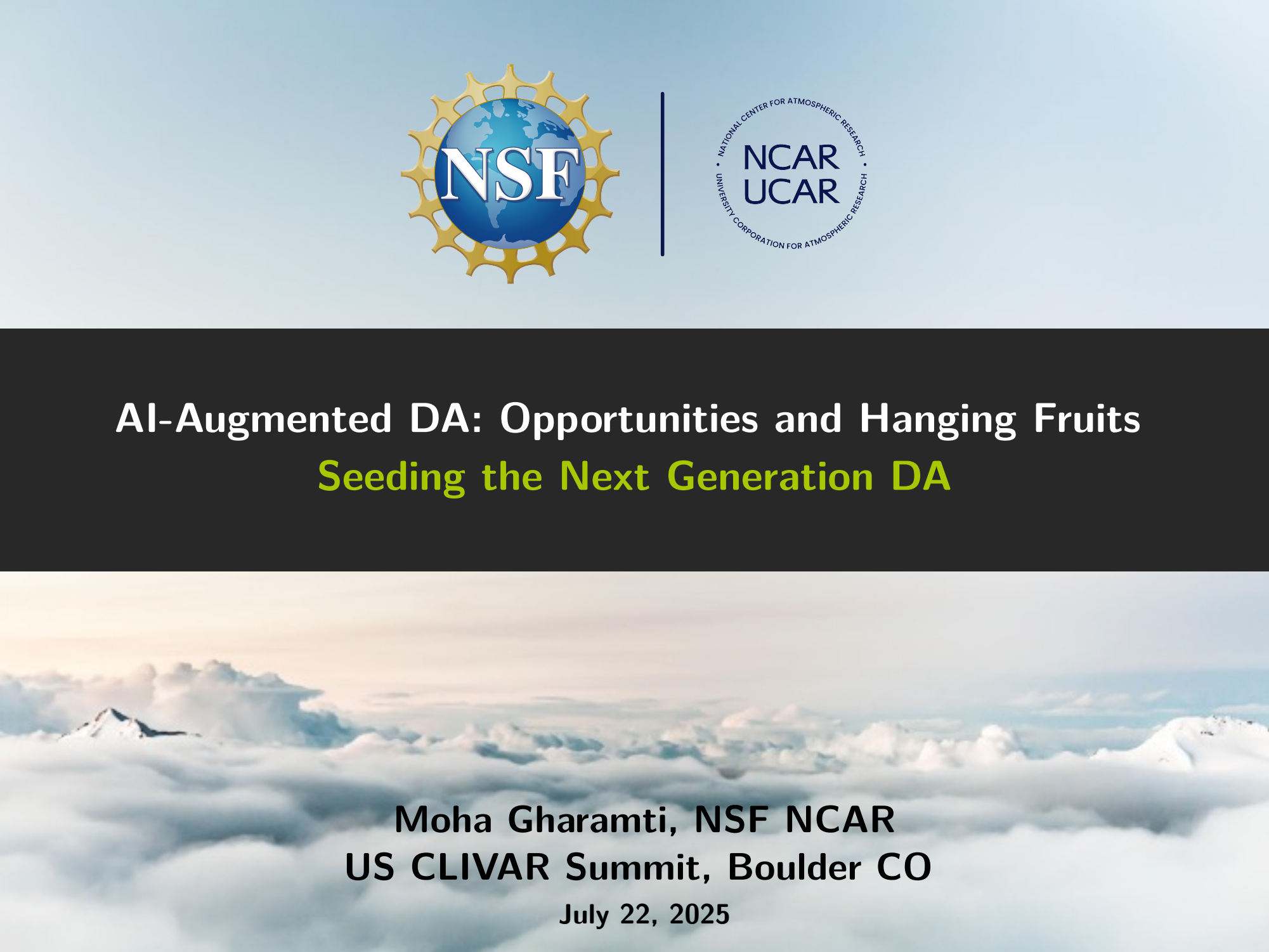


AI-Augmented DA: Opportunities and Hanging Fruits

Seeding the Next Generation DA

Moha Gharamti, NSF NCAR

US CLIVAR Summit, Boulder CO

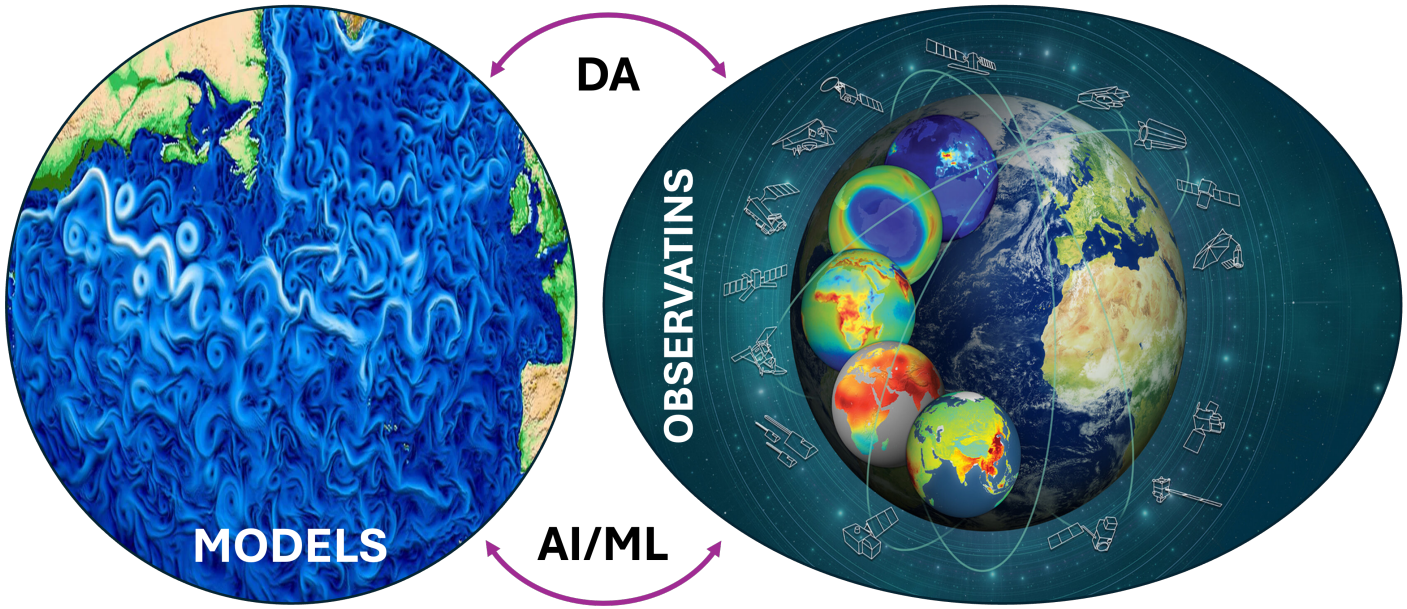
July 22, 2025

Motivation: Why AI in DA?

- In DA, we attempt to solve the following problem:

$$\underbrace{\left(\mathbf{x}-\mathbf{x}_b\right)^T \mathbf{B}^{-1}\left(\mathbf{x}-\mathbf{x}_b\right)}_{J_b\text{: background}} + \sum_{k=0}^{N} \underbrace{\left(\mathbf{y}_k-\mathbf{H}_k\left(\mathbf{x}\right)\right)^T \mathbf{R}_k^{-1}\left(\mathbf{y}_k-\mathbf{H}_k\left(\mathbf{x}\right)\right)}_{J_o\text{: Observation}}$$

- DA already integrates physics and data, so why not AI too?
- AI $\neq$ replacement, but rather a tool to **enhance** DA



The Hanging Fruits

1. **Accurate Background Covariance**
2. **Enhanced Observation Handling**
3. **Improved Efficiency**
4. **Better Algorithms**
5. **Coupled DA**

1. Accurate Background Covariance ..

AI for flow-dependent, adaptive, realistic background error models

- **Bias correction:** Use ML to learn systematic model errors over time
  - NN to track and remove seasonal biases in model forecasts

1. Accurate Background Covariance ..

AI for flow-dependent, adaptive, realistic background error models

- **Bias correction:** Use ML to learn systematic model errors over time
  - NN to track and remove seasonal biases in model forecasts
- **Flow-dependent covariances:** Learn spatial/temporal structures from past model-observation mismatches
  - Train on reanalysis data

1. Accurate Background Covariance ..

AI for flow-dependent, adaptive, realistic background error models

- **Bias correction:** Use ML to learn systematic model errors over time
  - NN to track and remove seasonal biases in model forecasts
- **Flow-dependent covariances:** Learn spatial/temporal structures from past model-observation mismatches
  - Train on reanalysis data
- **Adaptive tuning:** $\rho \circ (\lambda \cdot \mathbf{B})$ Learn optimal localization and inflation dynamically using RL

1. Accurate Background Covariance ..

AI for flow-dependent, adaptive, realistic background error models

- **Bias correction:** Use ML to learn systematic model errors over time
  - NN to track and remove seasonal biases in model forecasts
- **Flow-dependent covariances:** Learn spatial/temporal structures from past model-observation mismatches
  - Train on reanalysis data
- **Adaptive tuning:** $\rho \circ (\lambda \cdot \mathbf{B})$ Learn optimal localization and inflation dynamically using RL

$\Rightarrow$ AI/ML can help enrich what we know about uncertainty, not just states!

2. Enhanced Handling of Observations ..

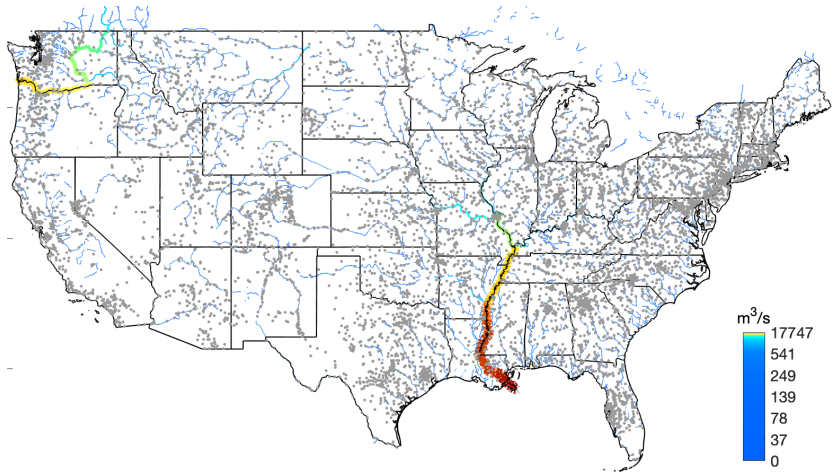
AI to expand, correct and reinterpret observations and their usage

- **Observation operator correction:** ML can learn forward operators
  - Instead of using complex (linearized) RTMs, use a trained NN that's fast and tailored to the system

2. Enhanced Handling of Observations ..

AI to expand, correct and reinterpret observations and their usage

- **Observation operator correction:** ML can learn forward operators
  - Instead of using complex (linearized) RTMs, use a trained NN that's fast and tailored to the system
- **Pseudo-observations:** Use generative AI (diffusion models, VAE) to synthesize credible data in sparse areas
  - Streamflow in un-gauged basins (Flooding), rainfall in conflict zones
  - *Challenge:* How to assign errors?



2. Enhanced Handling of Observations ..

AI to expand, correct and reinterpret observations and their usage

- **Observation operator correction:** ML can learn forward operators
  - Instead of using complex (linearized) RTMs, use a trained NN that's fast and tailored to the system
- **Pseudo-observations:** Use generative AI (diffusion models, VAE) to synthesize credible data in sparse areas
  - Streamflow in un-gauged basins (Flooding), rainfall in conflict zones
  - *Challenge:* How to assign errors?
- **QC and anomaly detection:** Use unsupervised learning to spot bad observations early

2. Enhanced Handling of Observations ..

AI to expand, correct and reinterpret observations and their usage

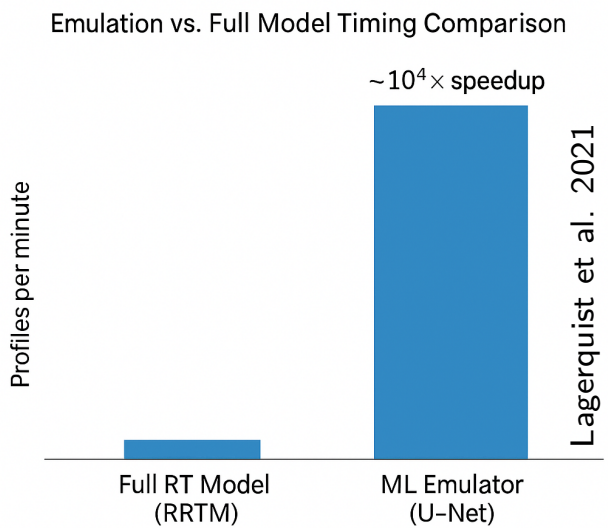
- **Observation operator correction:** ML can learn forward operators
  - Instead of using complex (linearized) RTMs, use a trained NN that's fast and tailored to the system
- **Pseudo-observations:** Use generative AI (diffusion models, VAE) to synthesize credible data in sparse areas
  - Streamflow in un-gauged basins (Flooding), rainfall in conflict zones
  - *Challenge:* How to assign errors?
- **QC and anomaly detection:** Use unsupervised learning to spot bad observations early

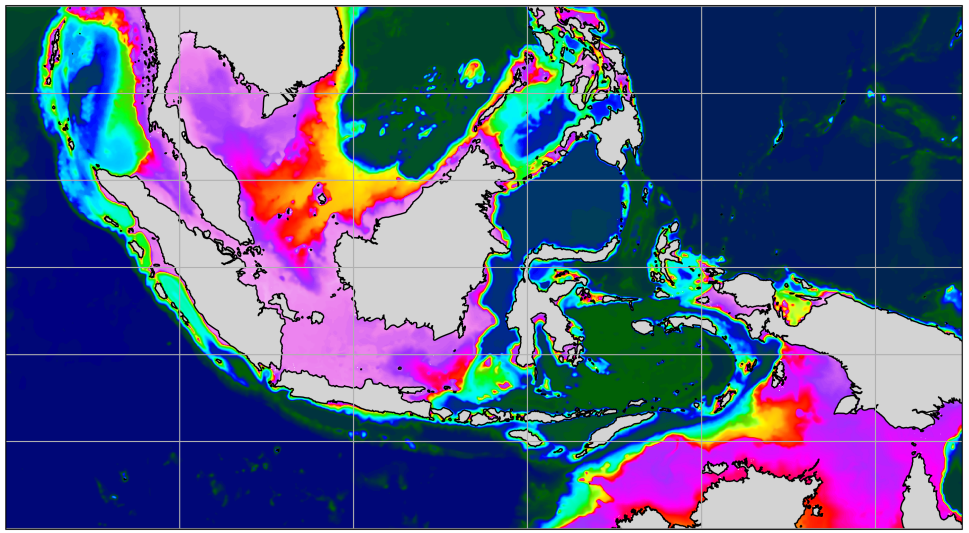
$\Rightarrow$ AI/ML can boost and refine observations, but we must understand uncertainty to use them!

3. Improved Efficiency ..

AI as shortcut or emulator for costly model components

- **Surrogates/emulators:** Emulate ocean BGC or cloud physics
  - Pre-train on offline data, use online in the DA loop

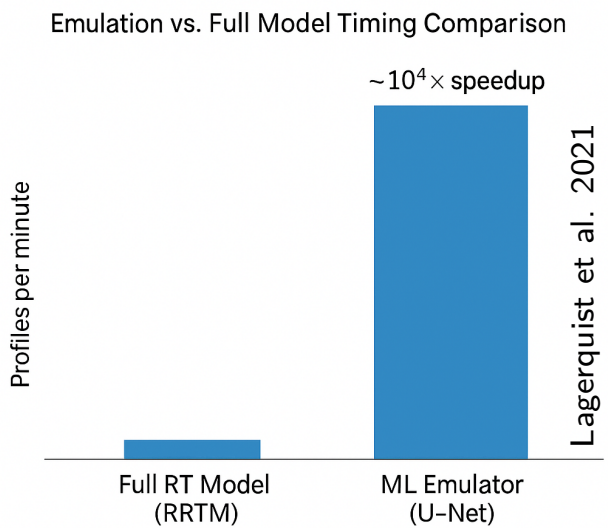


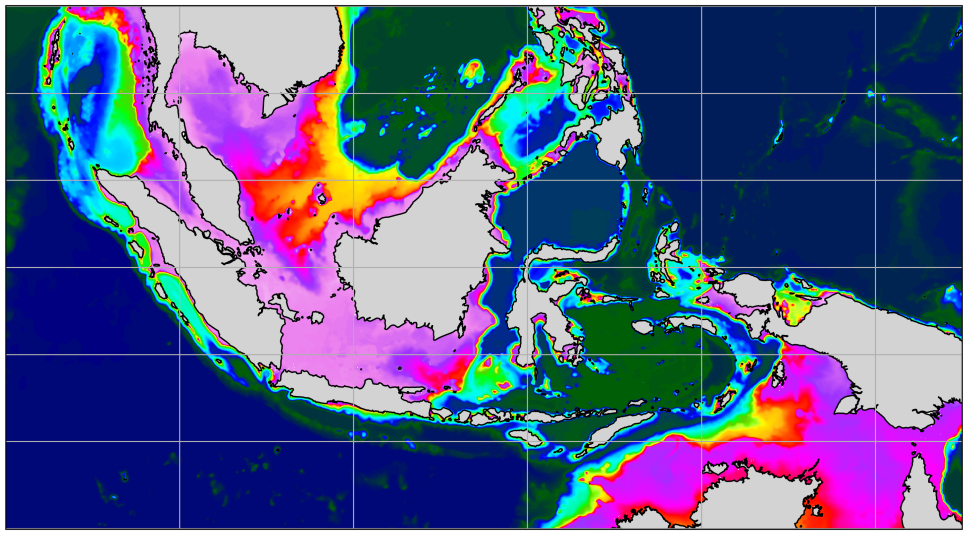


3. Improved Efficiency ..

AI as shortcut or emulator for costly model components

- **Surrogates/emulators:** Emulate ocean BGC or cloud physics
  - Pre-train on offline data, use online in the DA loop
- **Prior/Posterior sampling:** Capture important modes/structures of the state distribution
  - Use AI to re-weight or re-sample particles based on learned likelihoods
  - Ensure ensemble covers more realistic uncertainty





3. Improved Efficiency ..

AI as shortcut or emulator for costly model components

- **Surrogates/emulators:** Emulate ocean BGC or cloud physics
  - Pre-train on offline data, use online in the DA loop
- **Prior/Posterior sampling:** Capture important modes/structures of the state distribution
  - Use AI to re-weight or re-sample particles based on learned likelihoods
  - Ensure ensemble covers more realistic uncertainty

$\Rightarrow$ AI/ML can bring down computational cost without sacrificing accuracy!

4. Better Algorithms ..

AI-inspired methods to enhance the assimilation process itself

- **Hybrid DA-ML:** Combine physical and ML-based ensembles

$$\mathbf{B} = \alpha \mathbf{P}^{\text{ens}} + \beta \mathbf{P}^{\text{clim}} + \gamma \mathbf{P}^{\text{ML}},$$

$$\alpha + \beta + \gamma = 1$$

  - Errors-of-the-day using the flow-dependent ensemble
  - Long term variability using the climatology
  - Short-term biases with the ML ensemble

4. Better Algorithms ..

AI-inspired methods to enhance the assimilation process itself

- **Hybrid DA-ML:** Combine physical and ML-based ensembles

$$\mathbf{B} = \alpha \mathbf{P}^{\text{ens}} + \beta \mathbf{P}^{\text{clim}} + \gamma \mathbf{P}^{\text{ML}},$$

$$\alpha + \beta + \gamma = 1$$

  - Errors-of-the-day using the flow-dependent ensemble
  - Long term variability using the climatology
  - Short-term biases with the ML ensemble
- **Non-Gaussian DA:** Use normalizing flows for transformation
  - Helps tackle heavy tails, skewness

4. Better Algorithms ..

AI-inspired methods to enhance the assimilation process itself

- **Hybrid DA-ML:** Combine physical and ML-based ensembles

$$\mathbf{B} = \alpha \mathbf{P}^{\text{ens}} + \beta \mathbf{P}^{\text{clim}} + \gamma \mathbf{P}^{\text{ML}},$$

$$\alpha + \beta + \gamma = 1$$

  - Errors-of-the-day using the flow-dependent ensemble
  - Long term variability using the climatology
  - Short-term biases with the ML ensemble
- **Non-Gaussian DA:** Use normalizing flows for transformation
  - Helps tackle heavy tails, skewness
- **Parameter estimation:** Learn complex mappings between observations and model parameters (e.g., soil properties, turbulence)
  - Utilize Bayesian NNs to impose prior knowledge and uncertainty

4. Better Algorithms ..

AI-inspired methods to enhance the assimilation process itself

- **Hybrid DA-ML:** Combine physical and ML-based ensembles

$$\mathbf{B} = \alpha \mathbf{P}^{\text{ens}} + \beta \mathbf{P}^{\text{clim}} + \gamma \mathbf{P}^{\text{ML}},$$

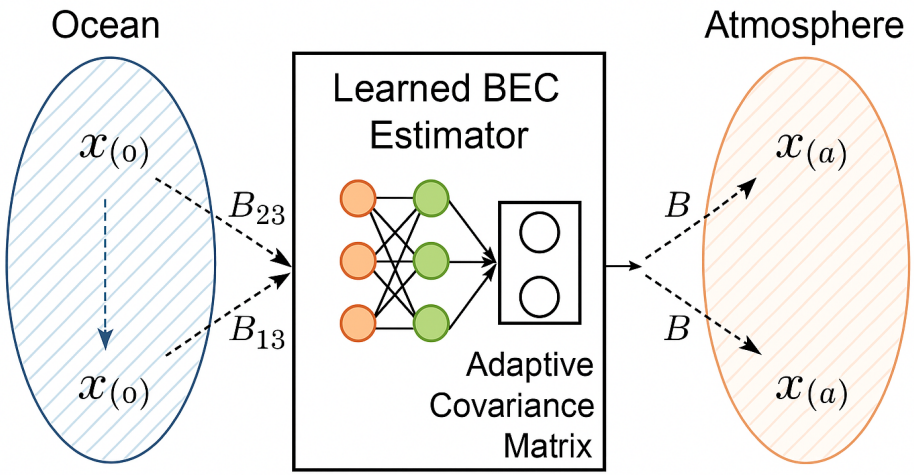
$$\alpha + \beta + \gamma = 1$$

  - Errors-of-the-day using the flow-dependent ensemble
  - Long term variability using the climatology
  - Short-term biases with the ML ensemble
- **Non-Gaussian DA:** Use normalizing flows for transformation
  - Helps tackle heavy tails, skewness
- **Parameter estimation:** Learn complex mappings between observations and model parameters (e.g., soil properties, turbulence)
  - Utilize Bayesian NNs to impose prior knowledge and uncertainty

$\Rightarrow$ AI/ML offers a bridge between theoretical advances (non-Gaussian, nonlinearity) and practical DA systems

5. Coupled DA ..

1. **Learning cross-component Covariances:** Train NNs on coupled reanalyses to learn mapping between different domains
2. **Data-driven Localization:** Use RL or supervised ML to adaptively select localization radii, especially at the interface
3. **Surrogate Cross-Covariance Estimators:** Use generative models to sample joint posterior distributions across components, capturing nonlinearity
4. **Regime-Aware Covariance Modeling:** Use classification/clustering to identify distinct dynamical regimes (e.g., ENSO, MJO phases). Switch/blend covariance structures accordingly



Cross-Cutting Questions

Some **big-picture thinking** is needed before boarding the DA+AI train:

- How do we handle uncertainty in AI-generated data?
- Can AI help where physics is poorly known or data are missing?
- How do we prevent **overfitting** when training AI on limited geophysical data?
- What **new metrics** are needed to evaluate AI-augmented DA?
- What role should human expertise play in supervising AI-augmented DA systems?
- How modular should AI components be in operational DA systems?
- Are there **theoretical limits** to what AI can learn about uncertainty?

Summary of Opportunities

Theme	AI Opportunity	Hanging Fruit
Background Covariance	Learn model bias, Flow-dependent BECs	Train bias estimators, RL for localization/inflation
Observations	Observation operators, Pseudo-observations, Automate QC	Generative-AI in sparse regions, Autoencoder-based QC
Efficiency	Emulate slow physical processes, Smarter prior/posterior sampling	Plug-in NN surrogates
Algorithms	Hybrid DA-ML systems, non-Gaussian transformations	Adaptive tuning, Use normalizing flows
Coupled DA	Learn cross-domain covariances and adaptive coupling behavior	Train on reanalyses, ML cross-covariances into EnKF

From AI-Augmented to AI-Native DA – How to get there?

Thank You!

gharamti@ucar.edu

The Data Assimilation Research Testbed (DART)

dart.ucar.edu dart@ucar.edu