Data Assimilation in Hydrology and Streamflow Forecasting

Hurricane Florence Flooding 2018

Moha Gharamti
https://dart.ucar.edu/
gharamti@ucar.edu

Special thanks to: James Mccreight, Tim Hoar, Seong Jin Noh, Arezoo RafieeiNasab, Ben Johnson, Nancy Collins

Date: June 3rd, 2021

National Center for Atmospheric Research
Data Assimilation Research Section (DAReS) - TDD - CISL

Overview

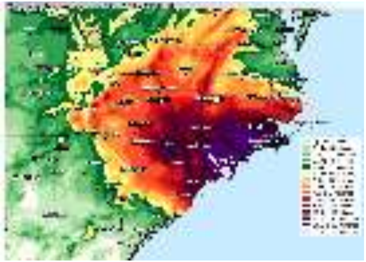
1. **Motivation**
2. **The Model: WRF-Hydro**
3. **DART: The Data Assimilation Research Testbed**
4. **Model & DA Configuration**
5. **Hydrological Assessment**

Motivation

1. Why Streamflow Forecasting?

Hurricane Florence (2018):

- Tropical wave ⇝ tropical storm ⇝ **Category 4 Hurricane**
- Landfall on Sep. 14 (Carolinas) with winds up to 150 mph
- Catastrophic damages to coastal communities [$25 billion]
- Flooding magnitude greatly exceeded the levels observed due to Hurricane Matthew (2016) and Floyd (1999) **combined**



Rainfall estimates from Hurricane Florence (*Source: NWS*)

Hurricane Florence eye during landfall (*Source: NWS*)

1. Why Streamflow Forecasting?

Hurricane Florence flooding and damages; near Swansboro, NC (*Source: CBS 17*)

1. Why Streamflow Forecasting?

- Predicting major floods during extreme rainfall events is crucial
  1. Save lives (~ 50 people died due to Florence Flooding)
  2. Limit damages (via advance warnings)
  3. Protect infrastructure

Flooded city of New Bern, NC

Hurricane-related Facts

- 2020 was the most active season: 12 storms hit the continental US

Hurricane-related Facts

- 2020 was the most active season: 12 storms hit the continental US
- Some of the most lethal consequences of hurricane season are not the storms but their aftermath: since 2017 at least 39 people have died following storms because of carbon monoxide poisoning from improperly used generators

Hurricane-related Facts

- 2020 was the most active season: 12 storms hit the continental US
- Some of the most lethal consequences of hurricane season are not the storms but their aftermath: since 2017 at least 39 people have died following storms because of carbon monoxide poisoning from improperly used generators
- 2021 Atlantic hurricane season officially begun last Tuesday

Hurricane-related Facts

- 2020 was the most active season: 12 storms hit the continental US
- Some of the most lethal consequences of hurricane season are not the storms but their aftermath: since 2017 at least 39 people have died following storms because of carbon monoxide poisoning from improperly used generators
- 2021 Atlantic hurricane season officially begun last Tuesday
- NHC have 21 storm names ready for this season:
  Ana, Bill, Claudette, Danny, Elsa, Fred, Grace, Henri, Ida, Julian, Kate, Larry, Mindy, Nicholas, Odette, Peter, Rose, Sam, Teresa, Victor and Wanda

The Model: WRF-Hydro

2.1 WRF-Hydro Objectives

WRF-Hydro: NCAR Weather Research and Forecasting model (WRF) hydrological modeling system. Research compartment of the **National Water Model (NWM)**.

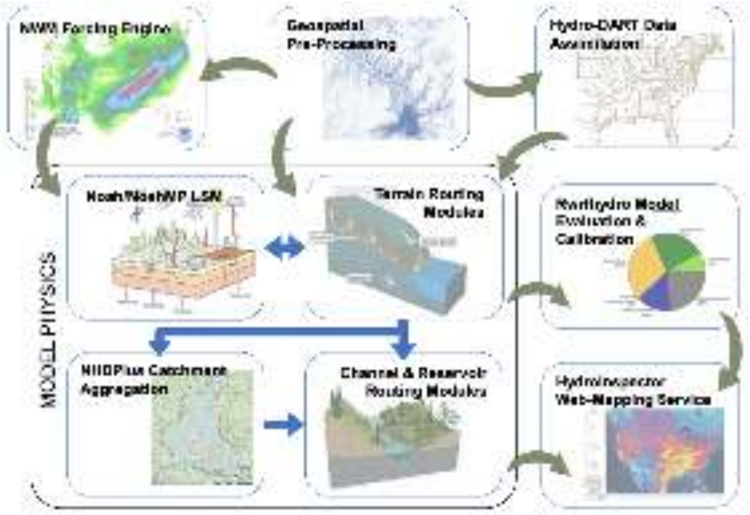
A community-based system, providing:

- Prediction of major water cycle components such as precipitation, soil moisture, snowpack, groundwater, streamflow, inundation
- Reliable streamflow prediction across scales (0-order headwater catchments to continental river basins and minutes to seasons)
- A robust framework for land-atmosphere coupling studies

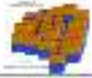
https://ral.ucar.edu/projects/wrf_hydro

Online Lessons, Jupyternotbook lessons and applications, online exercises, training on DockerHub, ...

2.2 Full WRF-Hydro Ecosystem



2.3 Full WRF-Hydro Physics Permutations

		WRF-Hydro Options	Current NWM Configuration
Column Land Surface Model		3 up-to-date column land models: Noah, Noah-MP (w/ built-in multi-physics options), Sac-HTET	Noah-MP
Overland Flow Module		3 overland routing schemes: diffusive wave, kinematic wave, direct basin aggregation	Diffusive wave
Lateral Subsurface Flow Module		2 subsurface routing schemes: Boussinesq shallow saturated flow, 3d Richards model	Boussinesq shallow saturated flow
Conceptual Baseflow Parameterizations		2 groundwater schemes: direct aggregation storage-release: pass-through or exponential model	Exponential model
Channel Routing Hydraulics		5 channel flow schemes: diffusive wave, kinematic wave, RAPID, custom-network Muskingum or Musk.-Cunge	Custom-network (NHDPlus) Muskingum-Cunge model
Lake/Reservoir Management		1 lake routing scheme: level-pool management	Level-pool management

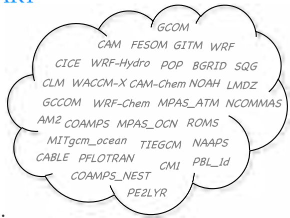
Streamflow (in cfs) simulation over CONUS for the 2019-2020 water year (*Source: NOAA, NWC, NWS*).

2.5 Streamflow Data

DART: The Data Assimilation Research Testbed

3.1 What is DART?

- A community facility for ensemble DA; developed and maintained by the Data Assimilation Research Section (DAReS) in CISL at NCAR
  - *Framework:*
    - Flexible, portable, well-tested, extensible, **free!**
    - Source code distributed on GitHub: NCAR/DART
    - Models: Toy to HUGE, including CESM
    - Observations: Real, synthetic, novel
  - *Research:*
    - Theory based, widely applicable techniques
    - Nonlinear filters, nonGaussian approaches
    - Adaptive inflation, Localization, ...
  - *Teaching:* Extensive tutorial materials and exercises



- ~ 50 UCAR member universities & more than 100 other sites
- Collaborations with external partners

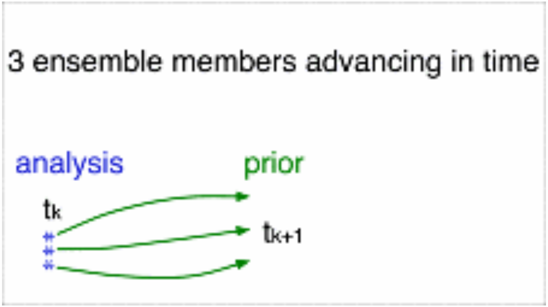
https://dart.ucar.edu/
https://docs.dart.ucar.edu/

3.2 Some DART Characteristics

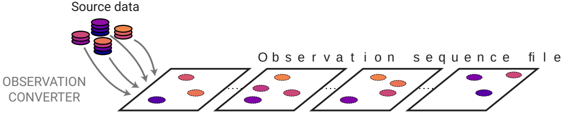
1. Assimilate the observations serially
   - remove the need to invert
   - simplify implementation, parallelism
   - equivalent to batch assimilation (localization usually breaks this)
2. Two-step least squares update scheme [Anderson 2003; MWR]
   - Find the observation increments; $\Delta y^{(i)} \quad i = 1, 2, \ldots, N_e$
   - Regress those increments in state space

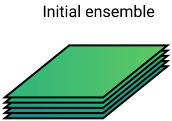
$$
\begin{aligned}
\Delta \mathbf{x}_j^{(i)} &= \sigma_{xy} \sigma_y^{-2} \Delta y^{(i)}, \\
\mathbf{x}_{j,k}^{a(i)} &= \mathbf{x}_{j,k}^{f(i)} + \alpha \Delta \mathbf{x}_j^{(i)}
\end{aligned}
$$

$$
\begin{aligned}
j &= 1, 2, \ldots, N_x \quad \text{(space)} \\
k &= 1, 2, \ldots, N_t \quad \text{(time)}
\end{aligned}
$$

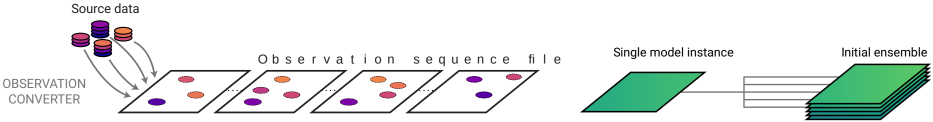


3.3 DART Flow & Functionality

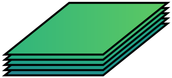




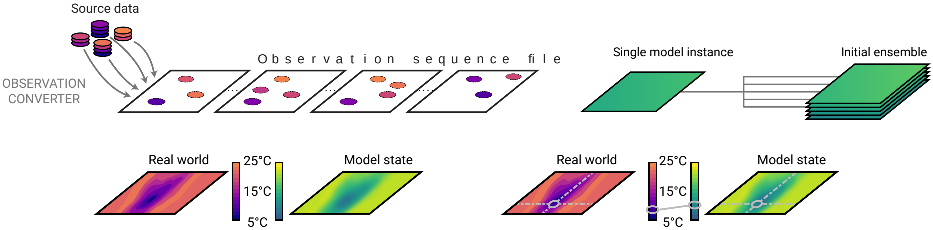
3.3 DART Flow & Functionality



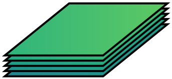
Initial ensemble



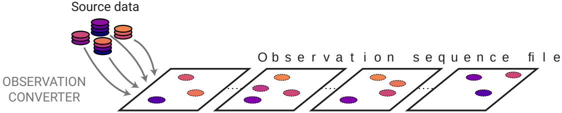
3.3 DART Flow & Functionality

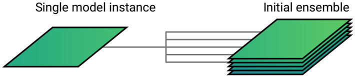


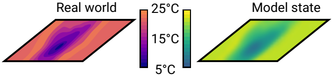
Initial ensemble

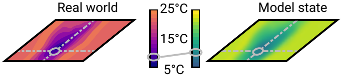


3.3 DART Flow & Functionality

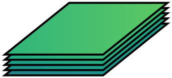




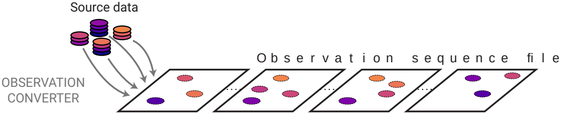


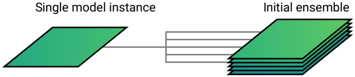


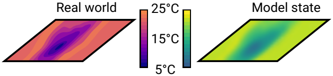
Initial ensemble

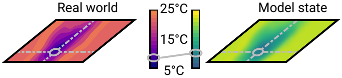


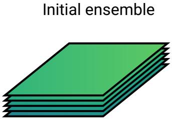
3.3 DART Flow & Functionality

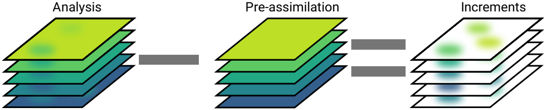








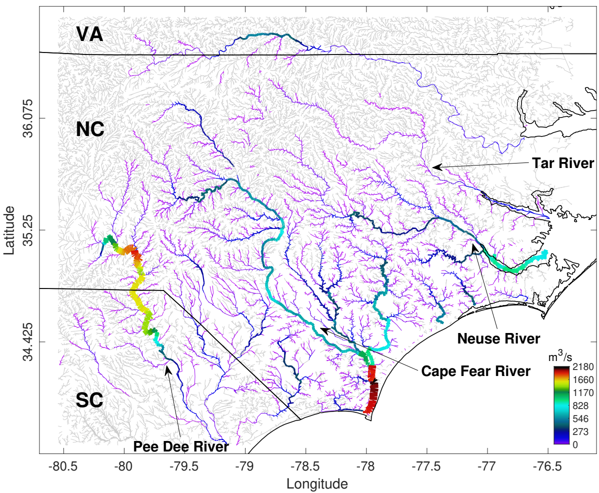




Model & DA Configuration

4.1 Model Domain and Observations

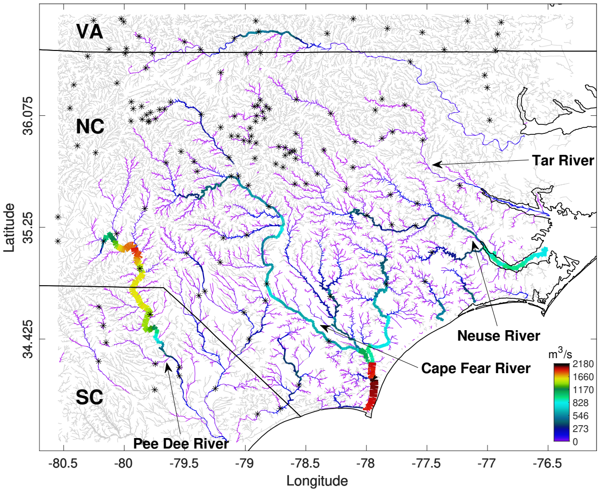
Interface DART [Anderson, 2008; BAMS] with WRF-Hydro (NOAA's NWM; Gochis, 2020) using HydroDART (refer to: NCAR/wrf_hydro_dart on GitHub)



- Regional subdomain of the NWM CONUS
- NWM channel network based on NHDPlus v.2
- ~ 67K reaches

4.1 Model Domain and Observations

Interface DART [Anderson, 2008; BAMS] with WRF-Hydro (NOAA's NWM; Gochis, 2020) using HydroDART (refer to: NCAR/wrf_hydro_dart on GitHub)



- Regional subdomain of the NWM CONUS
- NWM channel network based on NHDPlus v.2
- ~ 67K reaches
- Hourly streamflow assimilation
- 107 USGS gauges
- EAKF: 80 members

4.2 Spinup & DA Setup

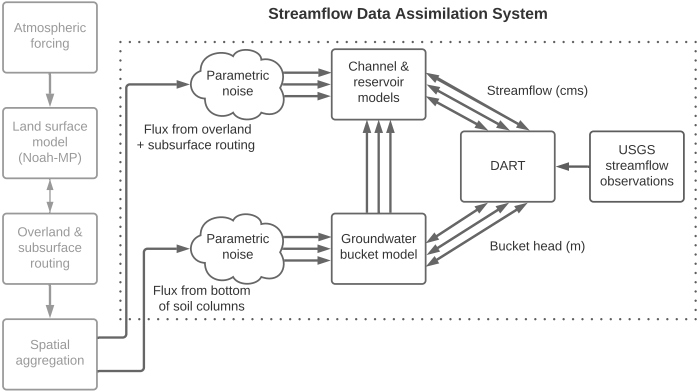
Channel + Bucket Configuration:

- Streamflow Model: Muskingum-Cunge hydrograph routing
- Groundwater Bucket Model: Mitigate baseflow deficincies

4.2 Spinup & DA Setup

Channel + Bucket Configuration:

- Streamflow Model: Muskingum-Cunge hydrograph routing
- Groundwater Bucket Model: Mitigate baseflow deficiencies



4.2 Spinup & DA Setup

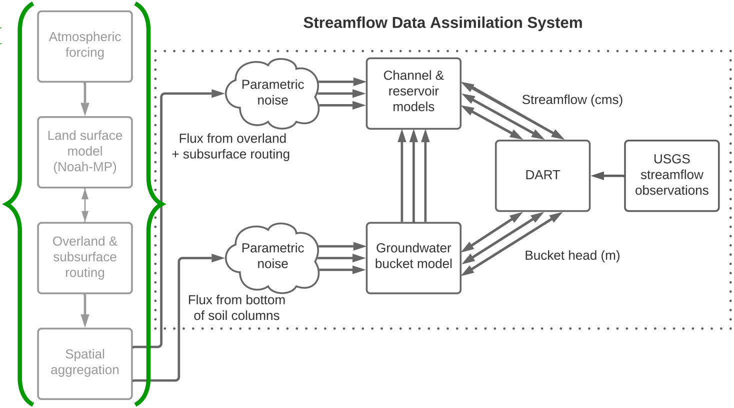
Channel + Bucket Configuration:

- Streamflow Model: Muskingum-Cunge hydrograph routing
- Groundwater Bucket Model: Mitigate baseflow deficiencies

Full model run from 2010-10-01 to 2018-07-01

beyond 2010-07-01: NWM operational analysis

Deterministic NWM model chain from forcing through aggregation



4.2 Spinup & DA Setup

Channel + Bucket Configuration:

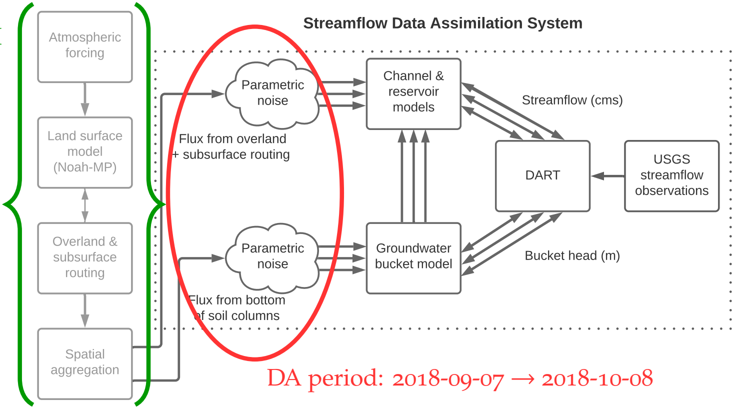
- **Streamflow Model:** Muskingum-Cunge hydrograph routing
- **Groundwater Bucket Model:** Mitigate baseflow deficincies

Full model run from 2010-10-01 to 2018-07-01

beyond 2010-07-01: NWM operational analysis

Deterministic NWM model chain from forcing through aggregation

One-way runoff fluxes used as input forcing to the channel+bucket sub-model



DA period: 2018-09-07 → 2018-10-08

4.3 Forcing and Ensemble Uncertainty

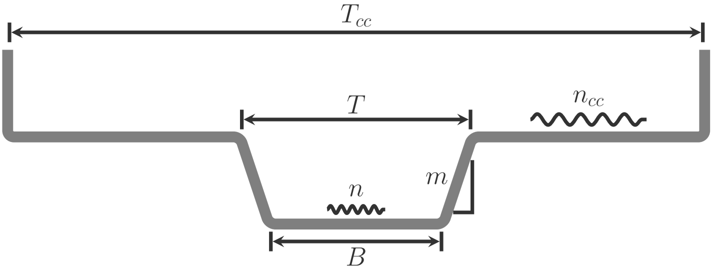
- Apply Gaussian perturbations to the boundary fluxes to the streamflow and bucket models every hourly forecast step

4.3 Forcing and Ensemble Uncertainty

- Apply Gaussian perturbations to the boundary fluxes to the streamflow and bucket models every hourly forecast step
- To create realistic model variability, we follow a "multi-configuration" approach and perturb the channel parameters:

1. top width, T
2. bottom width, B
3. side slope, m
4. Manning's N, n
5. width of compound channel, T_{cc}
6. Manning's N of compound channel, n_{cc}

Sampling uniformly under some physical constraints!



4.4.1 Along-The-Stream (ATS) Localization

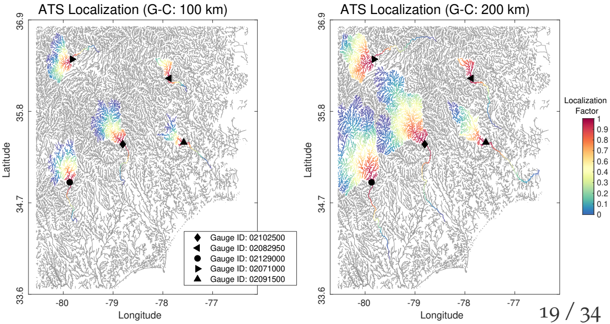
$$\mathbf{x}_{j,k}^{a(i)} = \mathbf{x}_{j,k}^{f(i)} + \alpha \Delta x_j^{(i)} \qquad 0 < \alpha < 1 \qquad \text{(Localization Factor)}$$

- Small ensemble sizes produce imperfect sample covariances [Houtekamer and Mitchell, 2001; MWR], yielding spurious correlations
- ATS localization [El Gharamti et al., 2020; HESS] aims to mitigate not only spurious correlations, due to limited ensemble size, but also **physically incorrect correlations between unconnected state variables in the river network**

4.4.1 Along-The-Stream (ATS) Localization

$$\mathbf{x}_{j,k}^{a(i)} = \mathbf{x}_{j,k}^{f(i)} + \alpha \Delta x_j^{(i)} \qquad 0 < \alpha < 1 \qquad (\text{Localization Factor})$$

- Small ensemble sizes produce imperfect sample covariances [Houtekamer and Mitchell, 2001; MWR], yielding spurious correlations
- ATS localization [El Gharamti et al., 2020; HESS] aims to mitigate not only spurious correlations, due to limited ensemble size, but also **physically incorrect correlations between unconnected state variables in the river network**

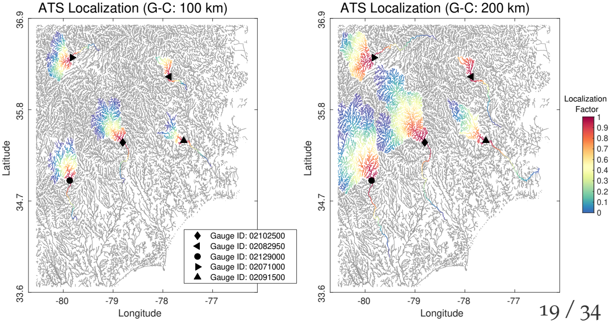


4.4.1 Along-The-Stream (ATS) Localization

$$\mathbf{x}_{j,k}^{a(i)} = \mathbf{x}_{j,k}^{f(i)} + \alpha \Delta x_j^{(i)} \qquad 0 < \alpha < 1 \qquad \text{(Localization Factor)}$$

Some Characteristics:

1. Flow of information only travels downstream (tree-like shapes)
2. Total number of close reaches depend on the size of the basin
3. Observations in different catchments do not have common close reaches

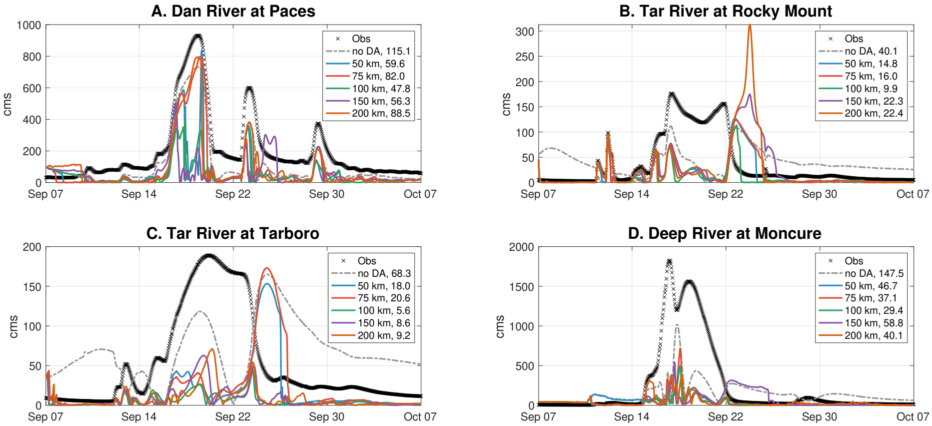


4.4.2 Does regular localization even work?

Tar River at Tarboro (NWIS 02083500)		ATS	Reg 20	Reg 10	Reg 5	Reg 2	Reg 1
	Prior RMSE	5.58	18.54	8.86	33.46	41.61	34.32
	Posterior RMSE	4.93	17.82	6.75	25.11	33.66	26.41
	Prior Bias	-1.13	-11.65	-1.71	-20.24	-18.09	-11.07
	Posterior Bias	-0.85	-11.41	-0.74	-20.37	-17.16	-10.01
	Prior Spread	1.20	3.29	2.80	10.90	10.84	9.54
	Posterior Spread	1.55	3.00	2.27	6.28	6.43	5.17

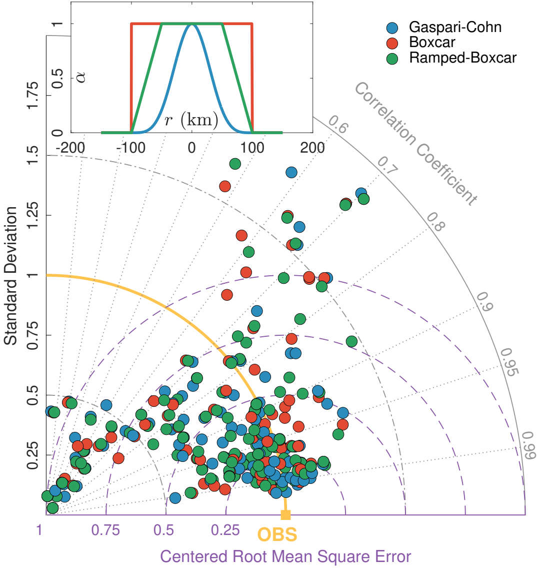
- Performance using ATS localization is significantly better ($\sim$ 40%)
- Using ATS, one can increase the effective localization radius
- Regular localization with large radii fails (correlating physically unrelated variables)

4.4.3 Tuning ATS Localization; [i] Radius



- Test with different localization radii: 50, 75, 100, 150, 200 km
- Larger radii degrade the accuracy (giving rise to spurious correlations)
- Smaller radii limit the amount of useful information
- Best performance with 100 km

4.4.4 Tuning ATS Localization; [ii] Correlation Function



- Averaging over all gauges, the correlation coefficient was: Gaspari-Cohn (**0.83**), Boxcar (**0.77**) and Ramped-Boxcar (**0.79**)
- Gaspari-Cohn outperforms other functions

4.5.1 Dealing with Variance Underestimation

- Variance underestimation often happens in ensemble-based systems due to sampling errors and model biases
- Other issues (that we usually ignore): High nonlinearity, nonGaussian features, correlation errors in the data

4.5.1 Dealing with Variance Underestimation

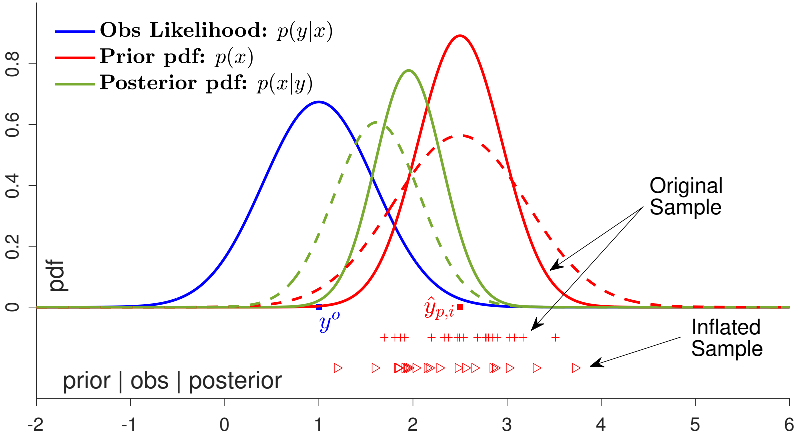
- Variance underestimation often happens in ensemble-based systems due to sampling errors and model biases
- Other issues (that we usually ignore): High nonlinearity, nonGaussian features, correlation errors in the data
- Inflation increases the variance around the ensemble mean:

$$\widetilde{\mathbf{x}}_j^{f|a(i)} \leftarrow \sqrt{\lambda}\left(\mathbf{x}_j^{f|a(i)} - \overline{\mathbf{x}}_j^{f|a}\right) + \overline{\mathbf{x}}_j^{f|a}$$

$f|a$ notation is used to refer either forecast or analysis. $\sqrt{\lambda}$ is the inflation factor. This scales the ensemble covariance by a λ:

$$\begin{aligned} \widetilde{\mathbf{P}}^{f|a} &= \lambda \cdot \mathbf{P}^{f|a} \\ &\equiv \lambda \sum_{i=1}^{N_e} \left(\mathbf{x}^{f|a(i)} - \overline{\mathbf{x}}^{f|a}\right)\left(\mathbf{x}^{f|a(i)} - \overline{\mathbf{x}}^{f|a}\right)^T \end{aligned}$$

4.5.1 Dealing with Variance Underestimation



4.5.2 How to choose $\sqrt{\lambda}$?

⋆ Spatially and Temporally Varying Adaptive Covariance Inflation [El Gharamti 2018; El Gharamti et al. 2019; MWR]:

1. Assume λ to be a random variable
2. Use the data to estimate λ at every point in the domain

4.5.2 How to choose $\sqrt{\lambda}$?

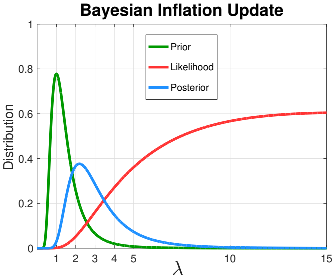
⋆ Spatially and Temporally Varying Adaptive Covariance Inflation [El Gharamti 2018; El Gharamti et al. 2019; MWR]:

1. Assume λ to be a random variable
2. Use the data to estimate λ at every point in the domain

Apply Bayes' rule:

$$p(\lambda|d) \approx p(\lambda) \cdot p(d|\lambda)$$

- **Prior** $p(\lambda)$; an Inverse Gamma pdf
- **Likelihood** $p(d|\lambda)$; a Gaussian function
  - $d = |y^o - \overline{\mathbf{x}}_j^f|$ is the innovation
  - Innovation statistics [Derosiers et al. 2005]:
    $\mathbb{E}(d) = 0; \quad \mathbb{E}(d^2) = \sigma_o^2 + \lambda\sigma_f^2$
- **Posterior** $p(\lambda|d)$



4.5.3 A quick illustration using DART_LAB's L96 GUIs

4.5.4 What to inflate; Prior or Posterior?

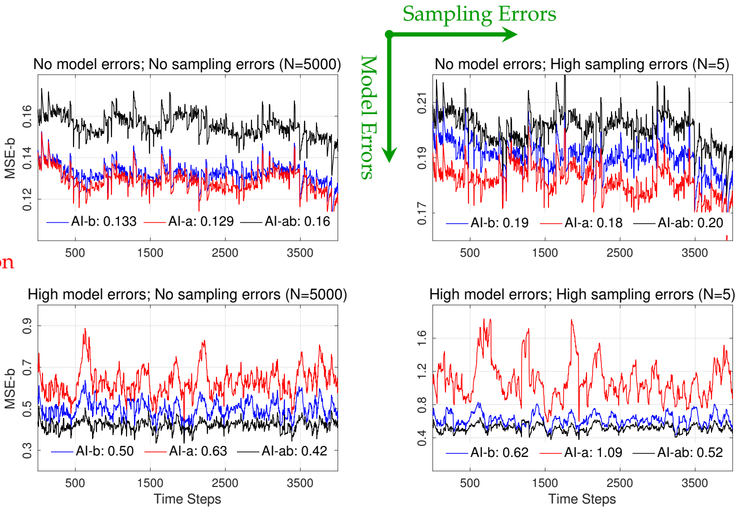
It is more common to inflate the prior covariance (after integrating the ensemble members forward in time). But what is more effective?

4.5.4 What to inflate; Prior or Posterior?

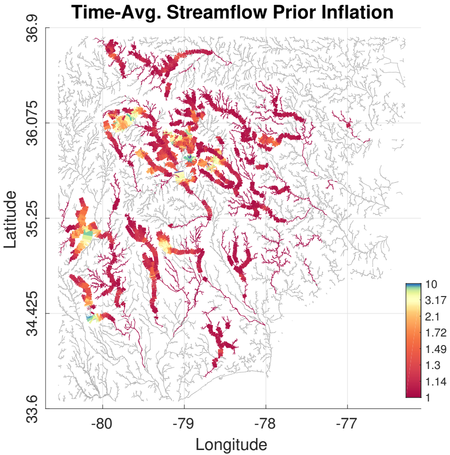
It is more common to inflate the prior covariance (after integrating the ensemble members forward in time). But what is more effective?

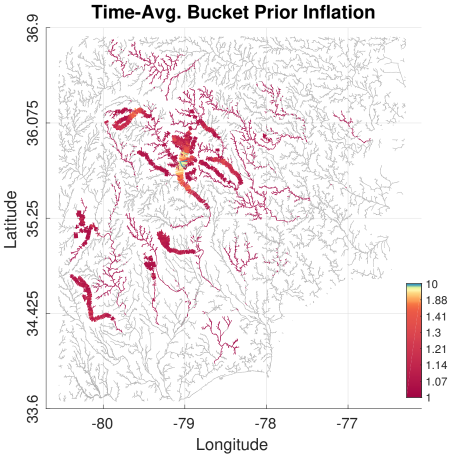
Lorenz-63 System

- ◇ prior inflation
- ◇ posterior inflation
- ◇ both



4.5.5 Inflation on the River Network

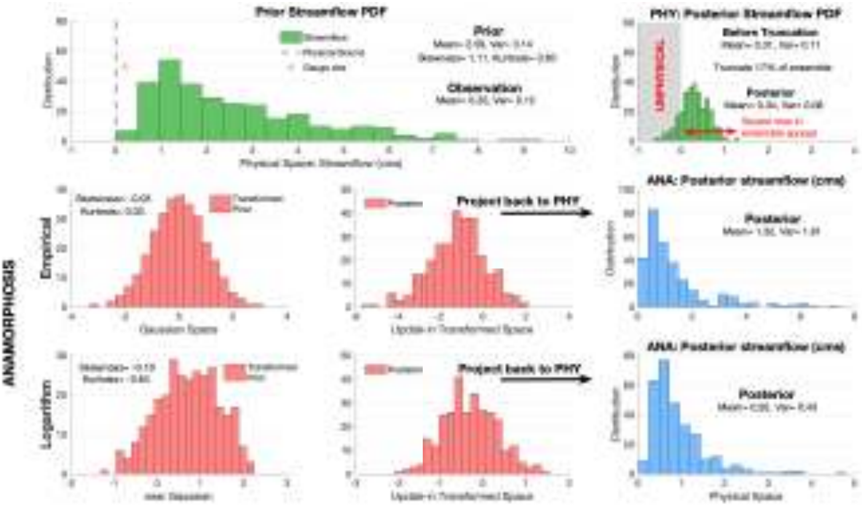




- Inflation follows tree-like shapes thanks to ATS localization
- Larger inflation in densely observed regions

4.6 Anamorphosis

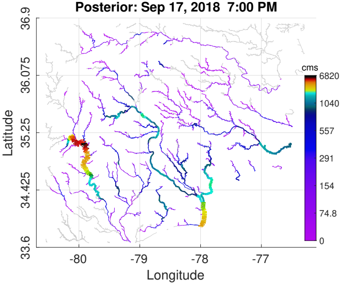
Streamflow is a positive quantity. We need to make sure the DA framework produces physically meaningful updates!

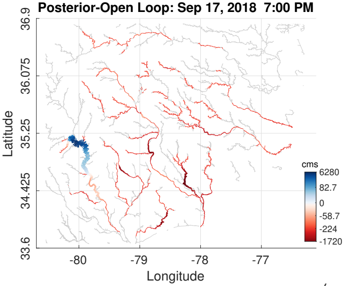


Hydrological Assessment

5.1 Bias Mitigation

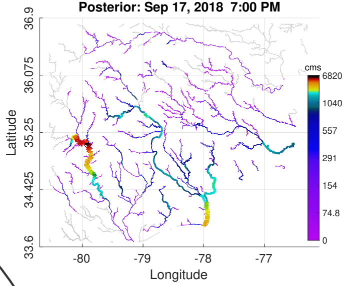
After landfall, the model's streamflow prediction (Open Loop) is significantly smaller than the posterior along Pee-Dee River in South Carolina

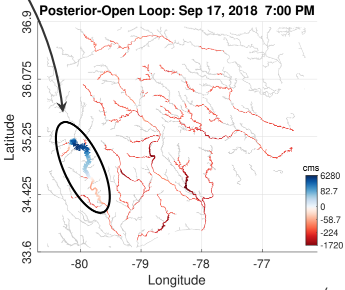




5.1 Bias Mitigation

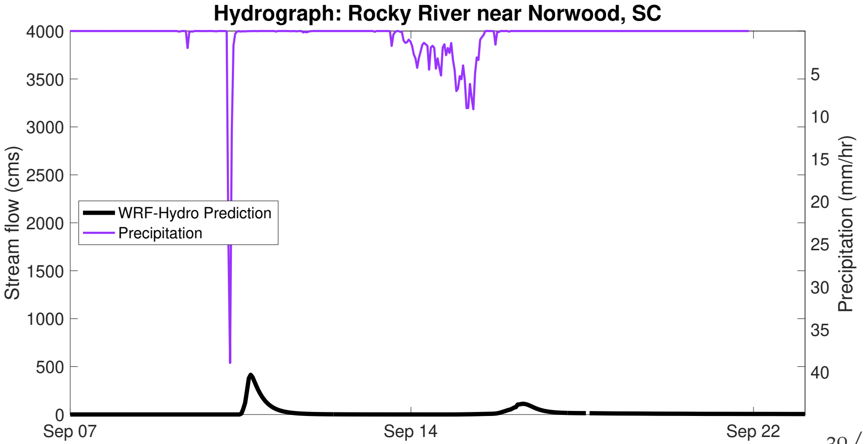
After landfall, the model's streamflow prediction (Open Loop) is significantly smaller than the posterior along Pee-Dee River in South Carolina





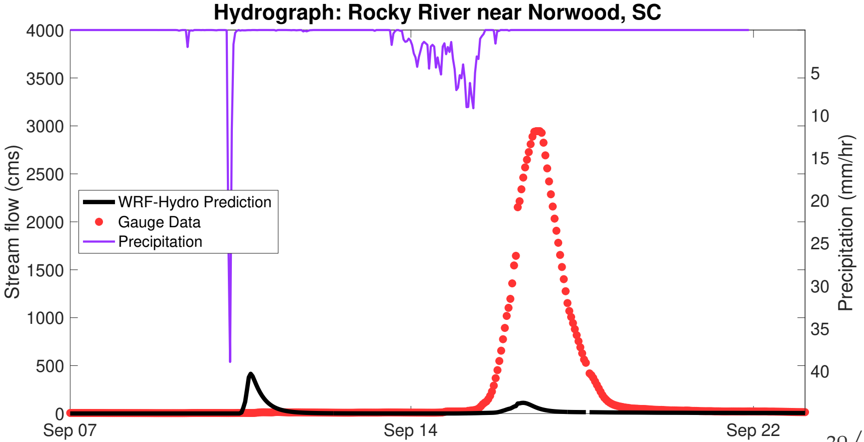
5.1 Bias Mitigation

After landfall, the model's streamflow prediction (Open Loop) is significantly smaller than the posterior along Pee-Dee River in South Carolina



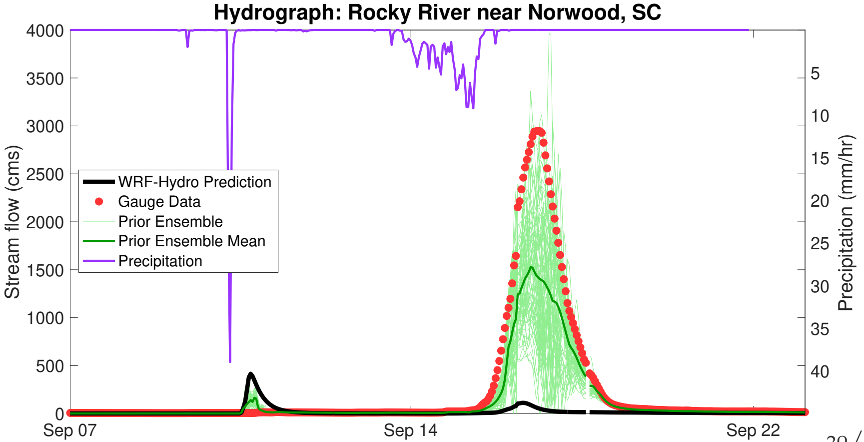
5.1 Bias Mitigation

After landfall, the model's streamflow prediction (Open Loop) is significantly smaller than the posterior along Pee-Dee River in South Carolina

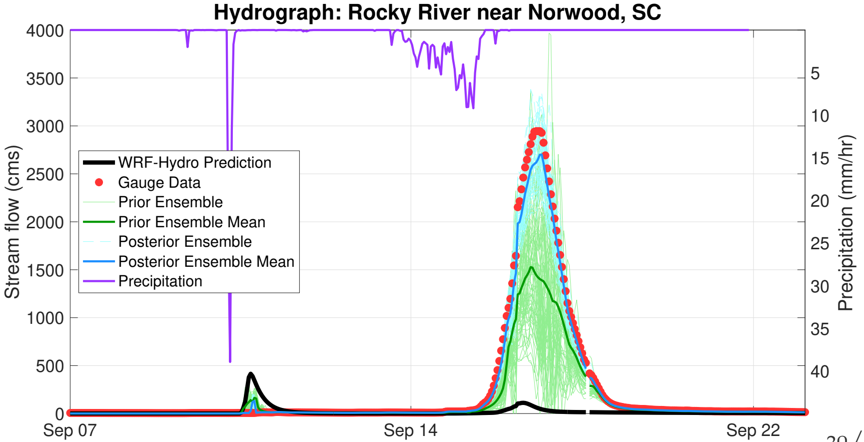


5.1 Bias Mitigation

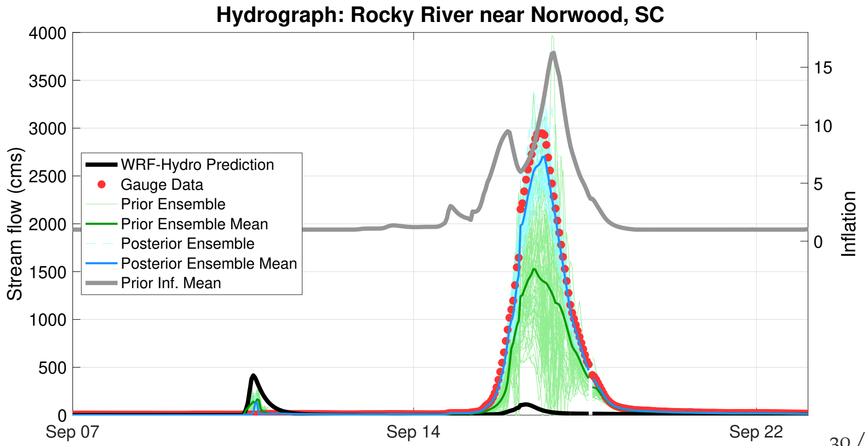
After landfall, the model's streamflow prediction (Open Loop) is significantly smaller than the posterior along Pee-Dee River in South Carolina



After landfall, the model's streamflow prediction (Open Loop) is significantly smaller than the posterior along Pee-Dee River in South Carolina

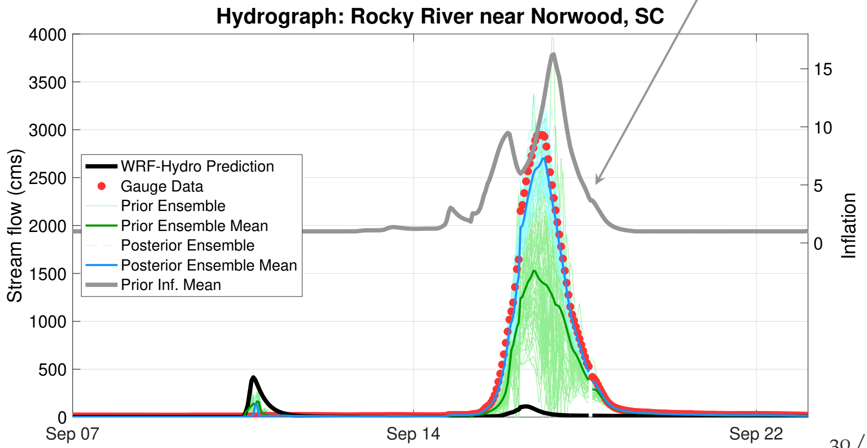


After landfall, the model's streamflow prediction (Open Loop) is significantly smaller than the posterior along Pee-Dee River in South Carolina



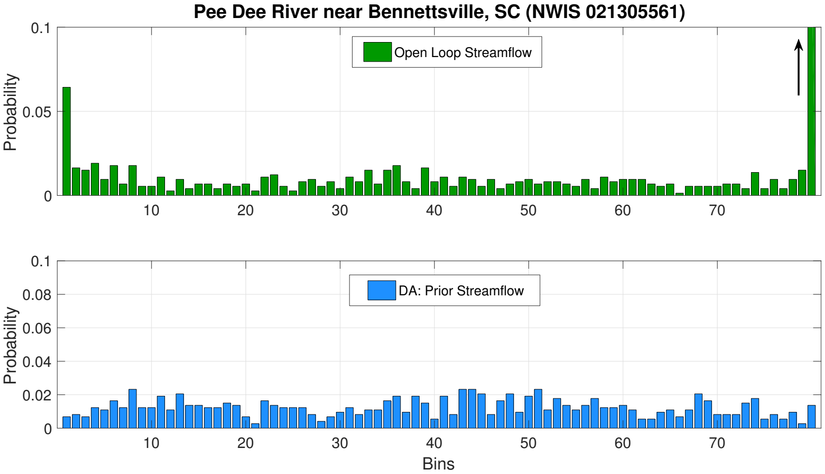
After landfall, the model's streamflow prediction (Open Loop) is significantly smaller than the posterior along Pee-Dee River in South Carolina

A sizable increase in prior inflation to counter the bias in the modeled streamflow!



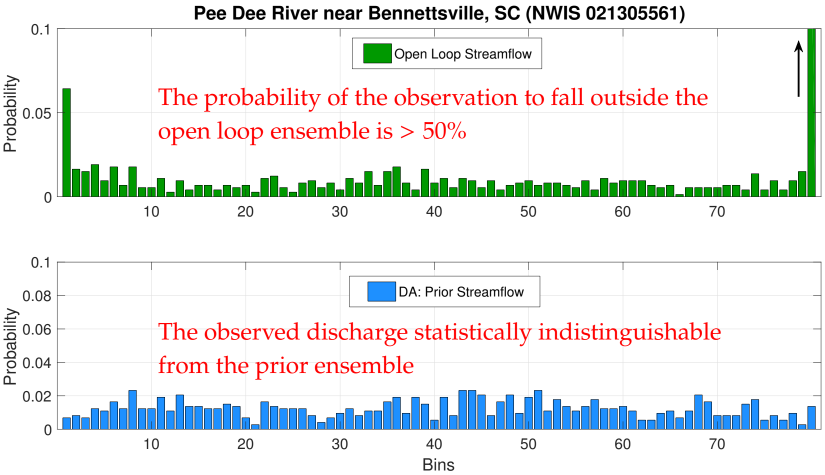
5.1 Bias Mitigation

The rank histogram for the open loop is heavily skewed to the right indicating that the gauge data is larger than the ensemble

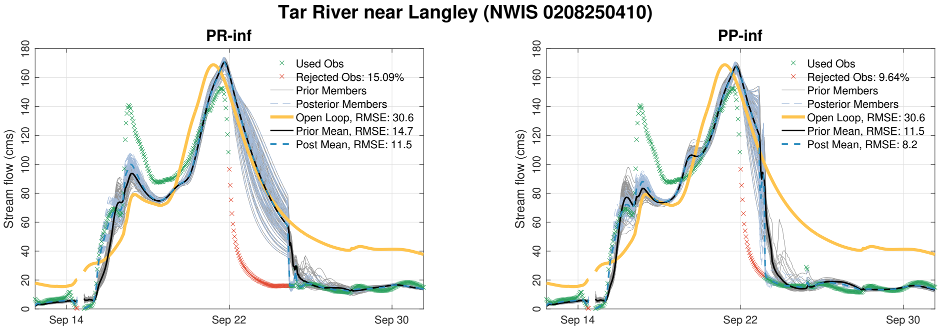


5.1 Bias Mitigation

The rank histogram for the open loop is heavily skewed to the right indicating that the gauge data is larger than the ensemble

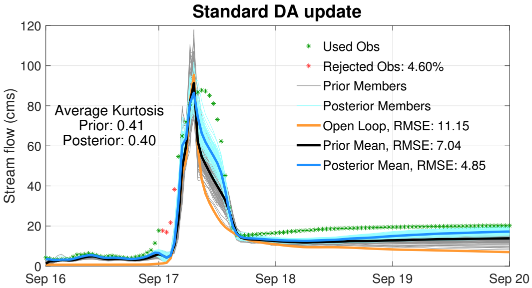


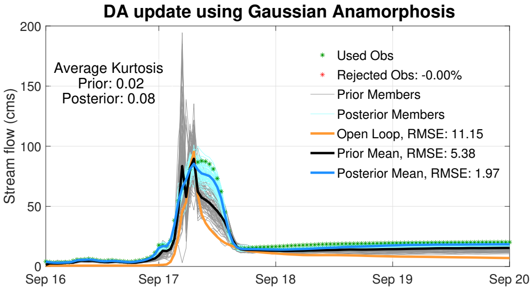
5.2 More on the effects of inflation



- Outlier Threshold: $|\overline{\mathbf{x}}_j^f - y^o| > \beta\sqrt{\sigma_o^2 + \sigma_f^2}$; $\beta = 3$
- Adding posterior inflation on top of prior inflation helps improve accuracy
- Falling limb of hydrograph (PP-inf) better fits the data. Recession happens almost 2 days earlier (rejects less data)
- May argue that posterior inflation could be resolving other regression issues such as sampling noise and nonGaussianity

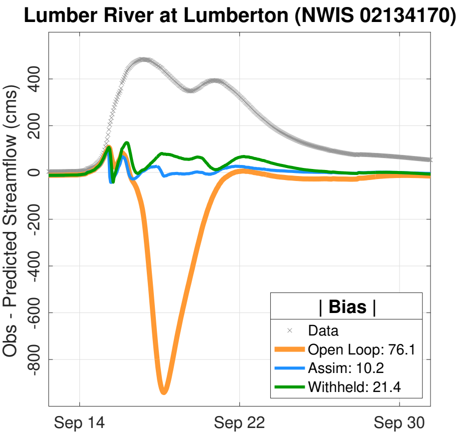
5.3 Benefits of Gaussian Anamorphosis

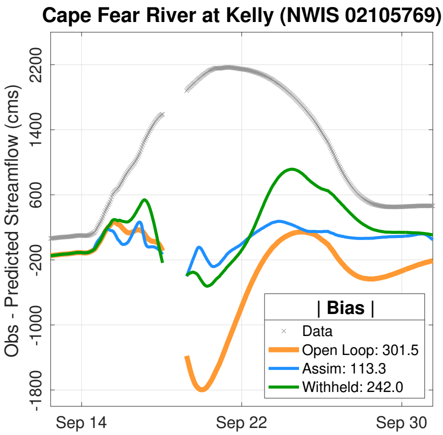




- Observation rejection is improved with GA
- Better fit to the observations on Sep. 17^{th}
- Higher order moments are almost completely eliminated using GA

5.4 Withholding Gauges

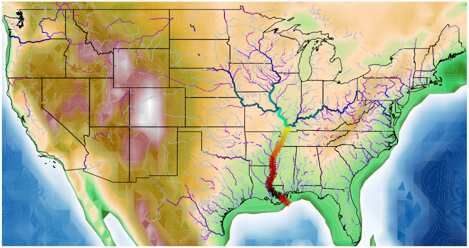




- By withholding gauges, we can infer the impact of the assimilation methods on un-gauged points within the domain
- DA is able to spread accurate information to unobserved locations

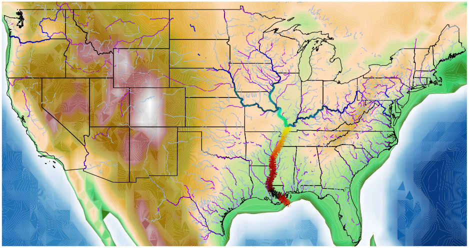
Future Research Directions

- Full CONUS streamflow reanalysis for the past 30 years: → Explore hybrid EnKF-OI approaches [El Gharamti 2021; MWR]



Future Research Directions

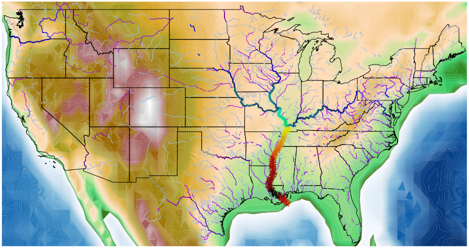
- Full CONUS streamflow reanalysis for the past 30 years:
  $\longrightarrow$ Explore hybrid EnKF-OI approaches
  [El Gharamti 2021; MWR]



- A collaborative project with USGS; 2 main goals:
  1. Assimilate gauge temperature data (investigate effects on streamflow)
  2. Placement of gauges (OSSE studies)

Future Research Directions

- Full CONUS streamflow reanalysis for the past 30 years:
  → Explore hybrid EnKF-OI approaches
  [El Gharamti 2021; MWR]



- A collaborative project with USGS; 2 main goals:
  1. Assimilate gauge temperature data (investigate effects on streamflow)
  2. Placement of gauges (OSSE studies)

- Coupling the LSM with WRF-Hydro:
  1. Assimilate soil moisture & streamflow; weak vs strong coupling
  2. Assimilate snow data (thickness, SWE, ...)