Mitigating the curse of small ensembles with

Probit-space Ensemble Size Expansion for Gaussian Copulas

(PESE-GC; "peace gee-see")

Man-Yau ("Joseph") Chan[1,2], Jeffrey L. Anderson[1] & Craig Schwartz[1]

[1]National Center for Atmospheric Research, USA

[2]The Ohio State University (starting Jan 2024), USA

Email: chan.1063@osu.edu

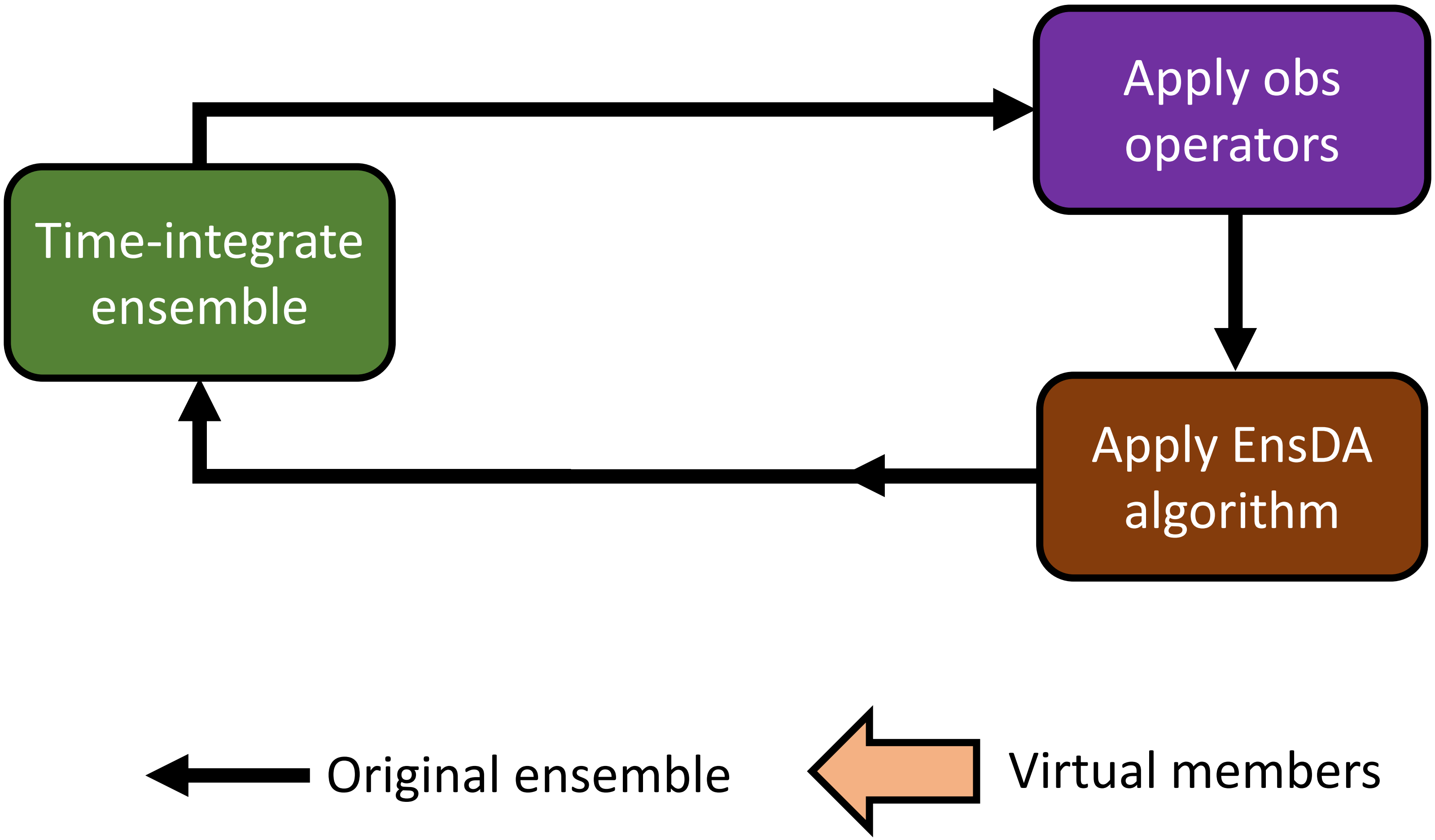
Motivation

Small ensemble sizes because models are expensive to run.

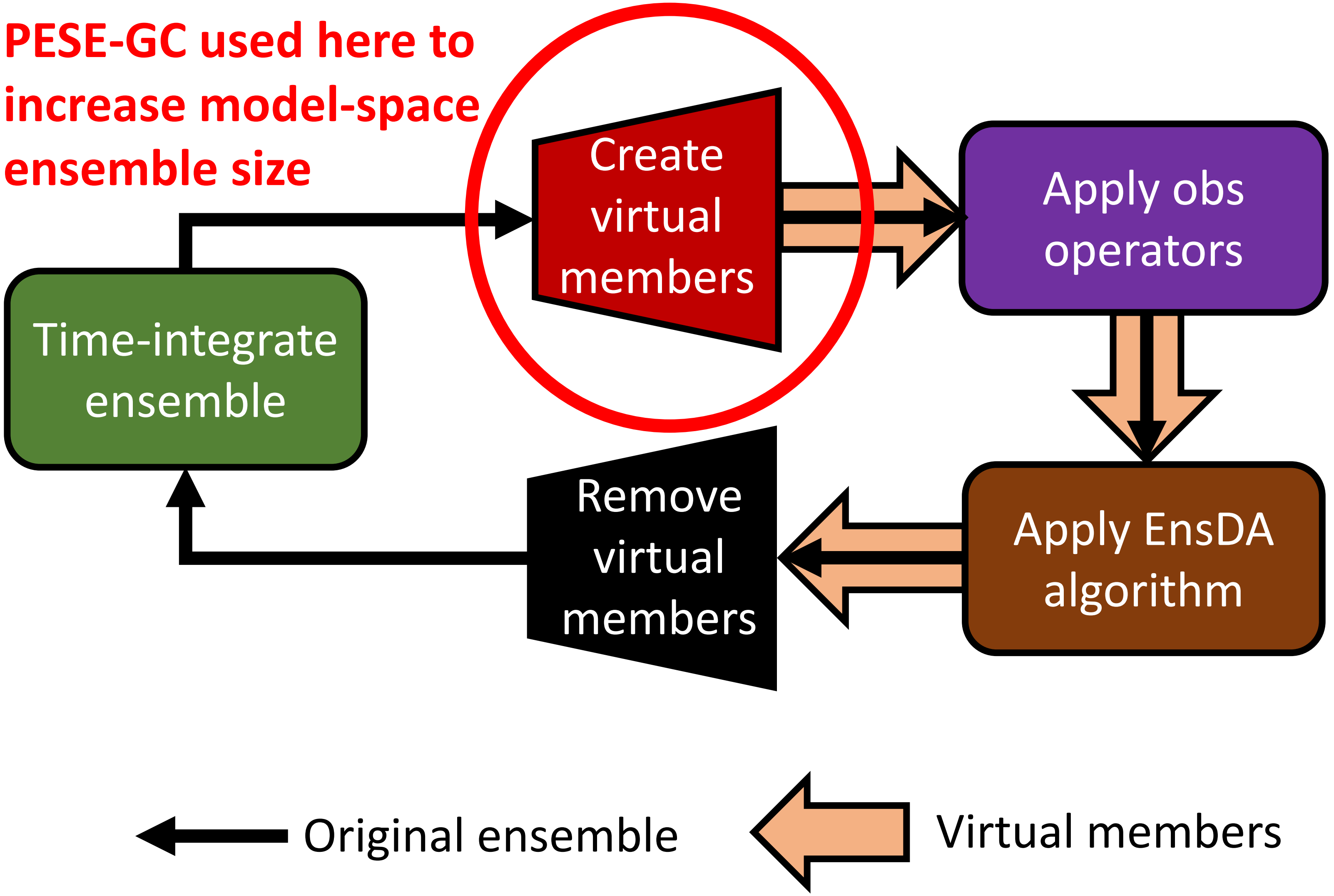
Thus, sampling errors contaminate ensemble statistics.

Therefore, limits EnsDA's impacts.

Goal: Improve EnsDA by increasing ensemble size without more model runs



Goal: Improve EnsDA by increasing ensemble size without more model runs



Question to ML folks in audience

What do you think of the usefulness of this method for ML in low-data situations?

Main Message

PESE-GC employs users’ knowledge of 1D forecast PDFs* to create additional model-space ensemble members.

This reduces sampling errors, thus improving EnsDA**.

* Aka, marginal forecast PDFs

** Tested using Lorenz 1996 model.

*Chan, Anderson, Chen (2020, Monthly Weather Review)

Starting point

Efficient and scalable Gaussian resampling algo*

<u>Traditional SVD recipe for Gaussian resampling:</u>

$$\boldsymbol{\psi_{virt}} = \boldsymbol{L} \; \boldsymbol{\psi} + \boldsymbol{\psi_{offset}}$$

Virtual member

A scarily large matrix

White noise vector

Making $\boldsymbol{L}$ via SVD requires $\sim(10^8)^3 = 10^{24}$ operations.

$\boldsymbol{L}$ must be computed online in DA executable.

Constructing $\boldsymbol{L}$ will likely cause parallelization bottlenecks.

*Chan, Anderson, Chen (2020, Monthly Weather Review)

Starting point

Efficient and scalable Gaussian resampling algo*

Fast weighted-sum recipe to make virtual members:

$$\boldsymbol{\psi_{virt}} = \sum_{n=1}^{N} \alpha_n \, \boldsymbol{\psi'_{old,n}} + \boldsymbol{\psi_{offset}}$$

Virtual member

Weighed sum of old perturbations

The weights (α_n) can be **determined offline in $\sim\mathbf{10^3}$ operations** and **model agnostic** (i.e., WRF & L96 use the same set of α_n).

This weighed sum procedure is $\sim\mathbf{10^{12}}$ **times more efficient** than traditional method.

Main Message

PESE-GC employs users' knowledge of 1D forecast PDFs* to create additional model-space ensemble members.

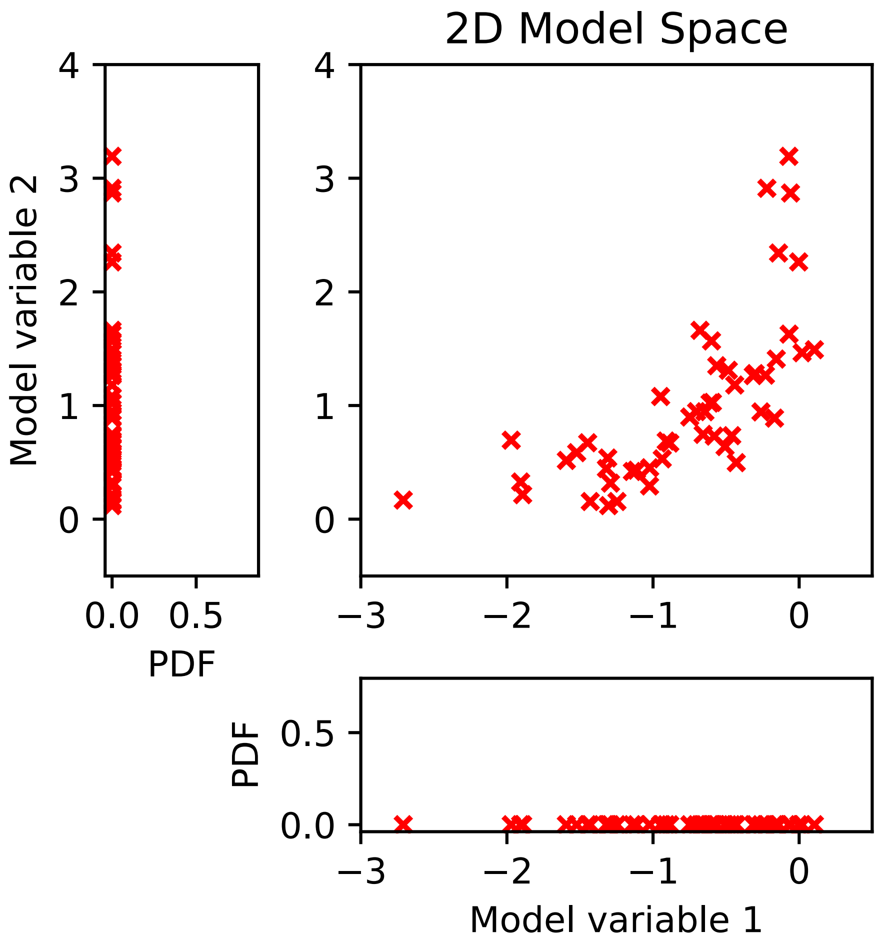
This reduces sampling errors, thus improving EnsDA**.

* Aka, marginal forecast PDFs

** Tested using Lorenz 1996 model.

PESE-GC's 4-step procedure

Employs users' knowledge of prior marginal PDFs & efficient resampling to generate virtual members.

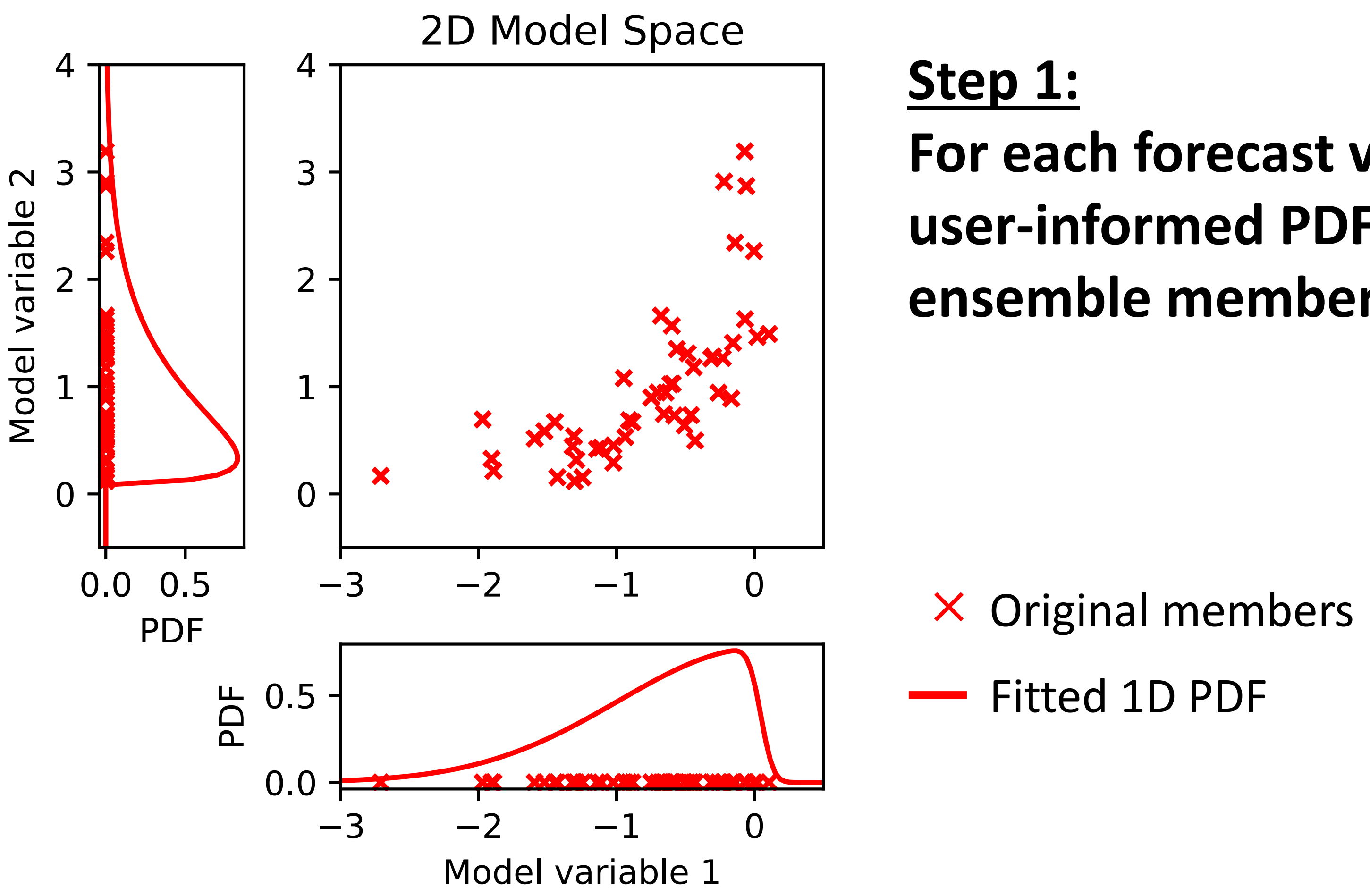


Will demo PESE-GC with a bivariate model space example.

× Original members

PESE-GC's 4-step procedure

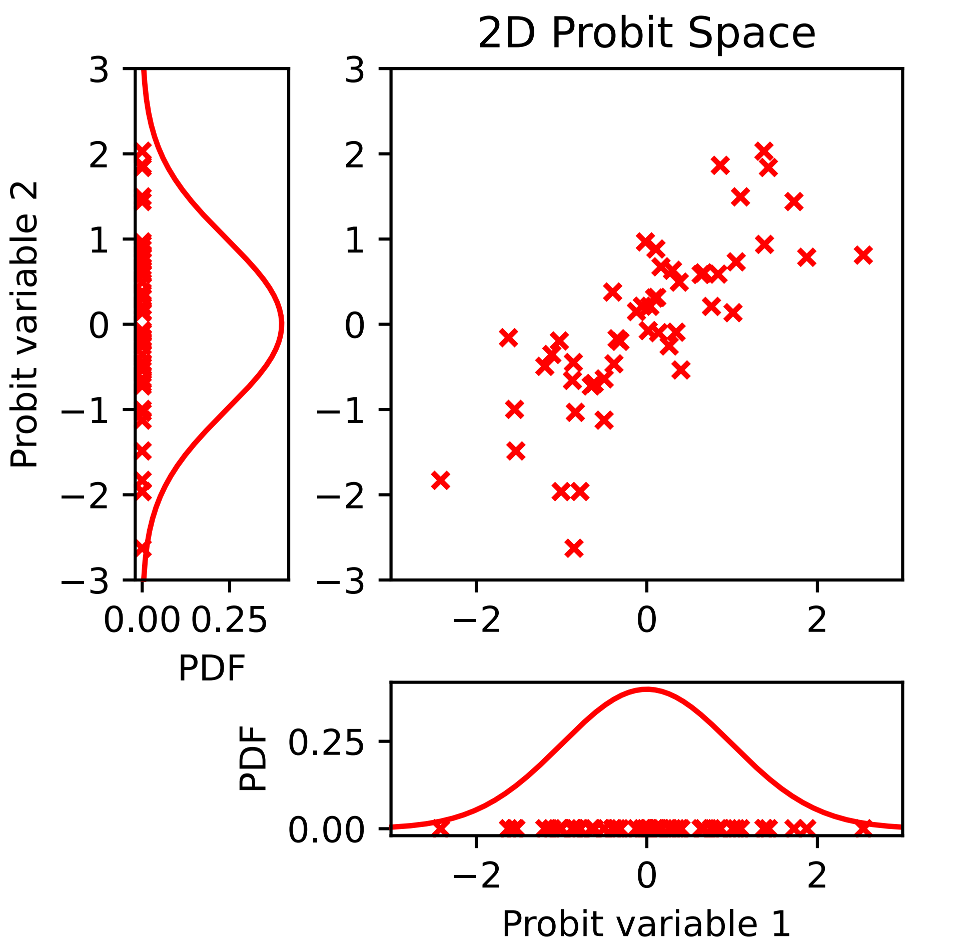
Employs users' knowledge of prior marginal PDFs & efficient resampling to generate virtual members.



<u>Step 1:</u>
For each forecast variable, fit user-informed PDF to original ensemble members

PESE-GC's 4-step procedure

Employs users' knowledge of prior marginal PDFs & efficient resampling to generate virtual members.



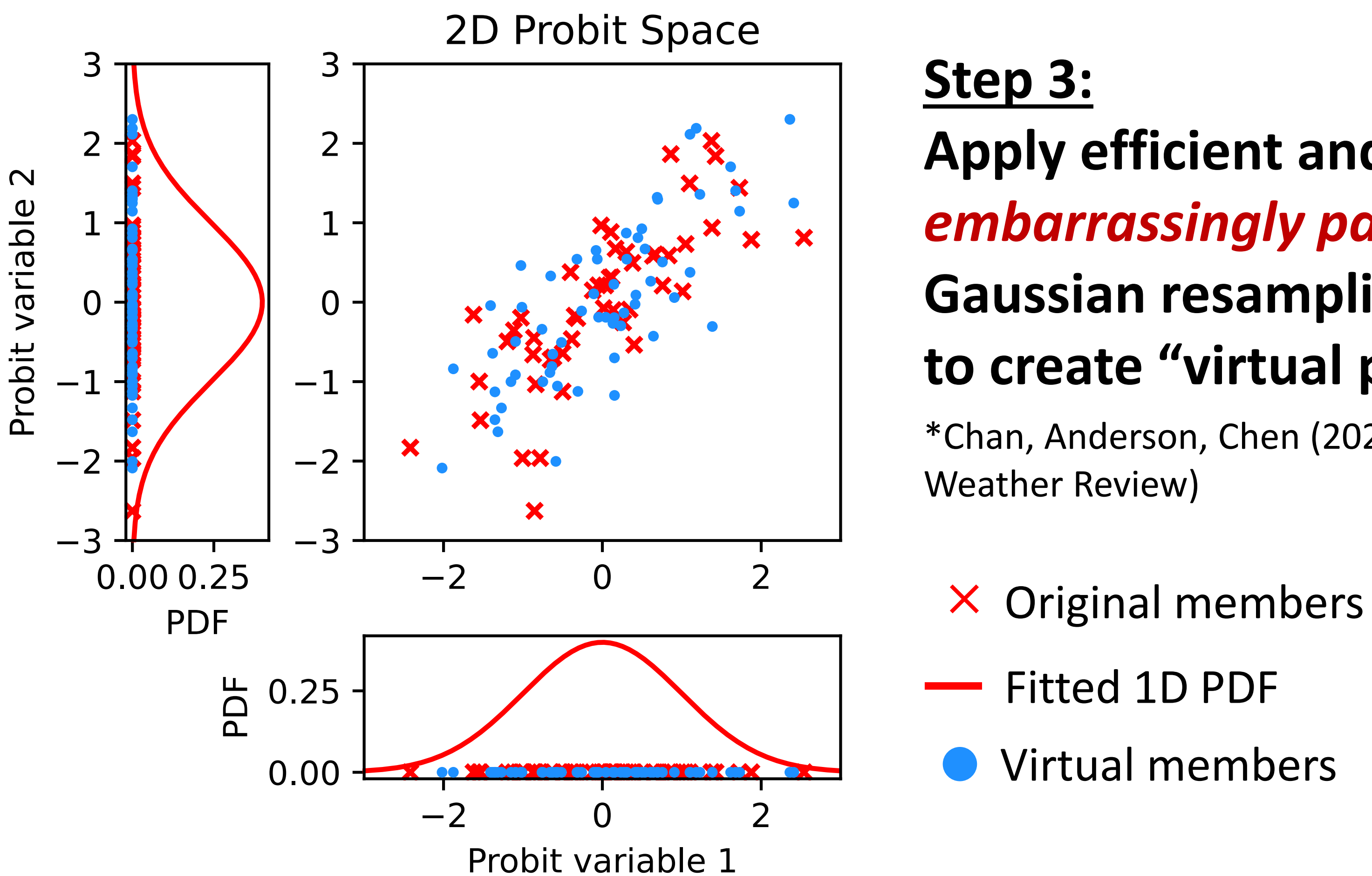
Step 2:

For each variable, transform fitted PDF and original members to standard normal.*

*A.k.a., probit probability integral transforms, a type of Gaussian anamorphosis

PESE-GC's 4-step procedure

Employs users' knowledge of prior marginal PDFs &
efficient resampling to generate virtual members.



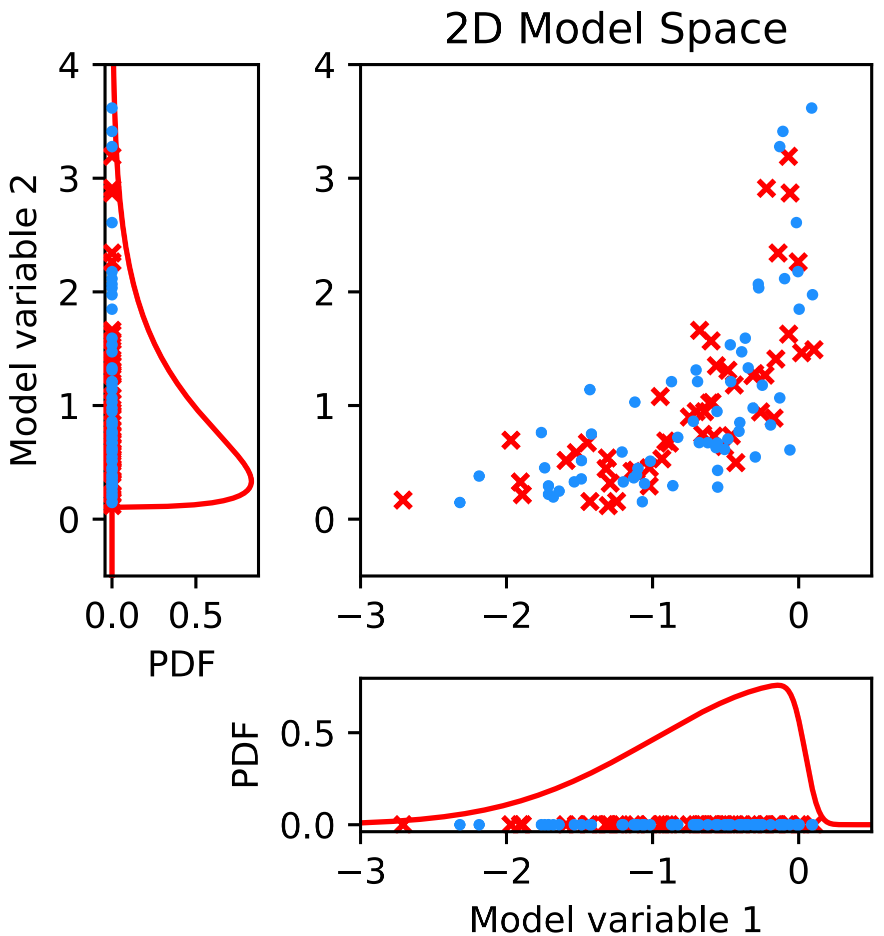
Step 3:

Apply efficient and *embarrassingly parallel* Gaussian resampling scheme* to create "virtual probits".

*Chan, Anderson, Chen (2020, Monthly Weather Review)

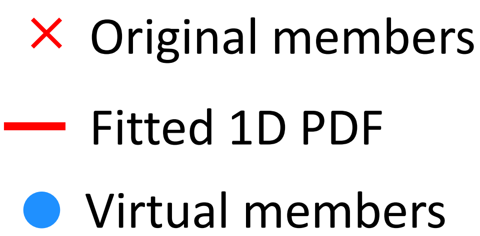
PESE-GC's 4-step procedure

Employs users' knowledge of prior marginal PDFs & efficient resampling to generate virtual members.



<u>Step 4:</u>

Reverse the transforms applied in step 2.



PESE-GC's 4 steps

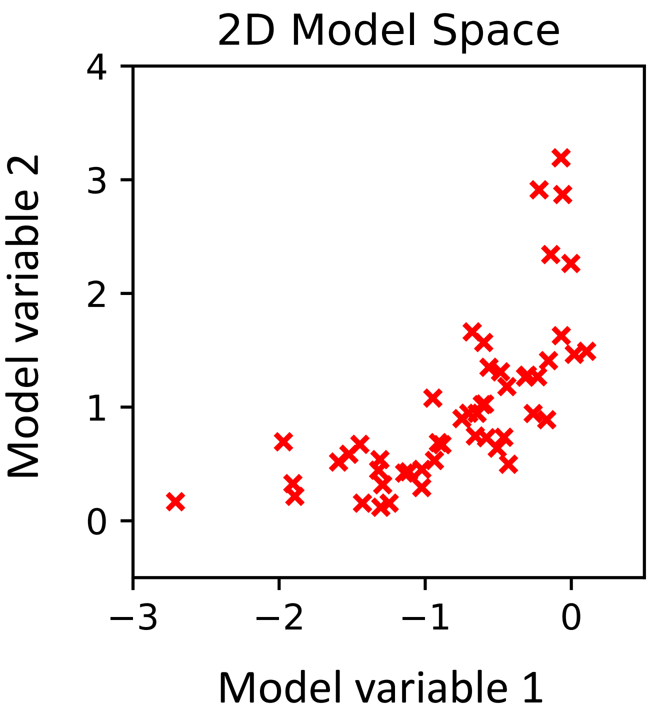
For each forecast model variable:

1. Fit user-informed marginal PDF to ensemble members.
2. Transform members into a Gaussian space ("Probit space").
3. Apply efficient and scalable Gaussian resampling.
4. Reverse transform applied in step 2.

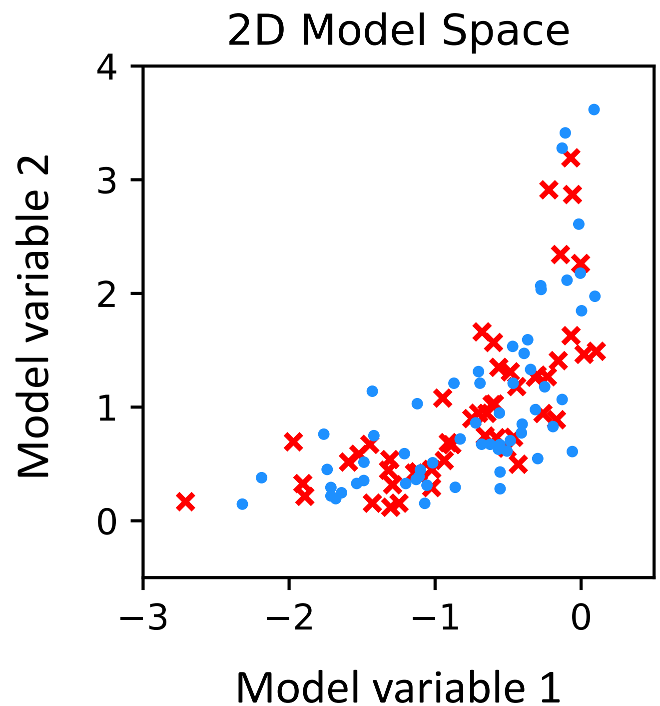
Note: ***These 4 steps are embarrassingly parallel!*** The execution speed scales well (linearly) with number of computer cores!

PESE-GC can handle non-Gaussian forecast distributions

Before PESE-GC

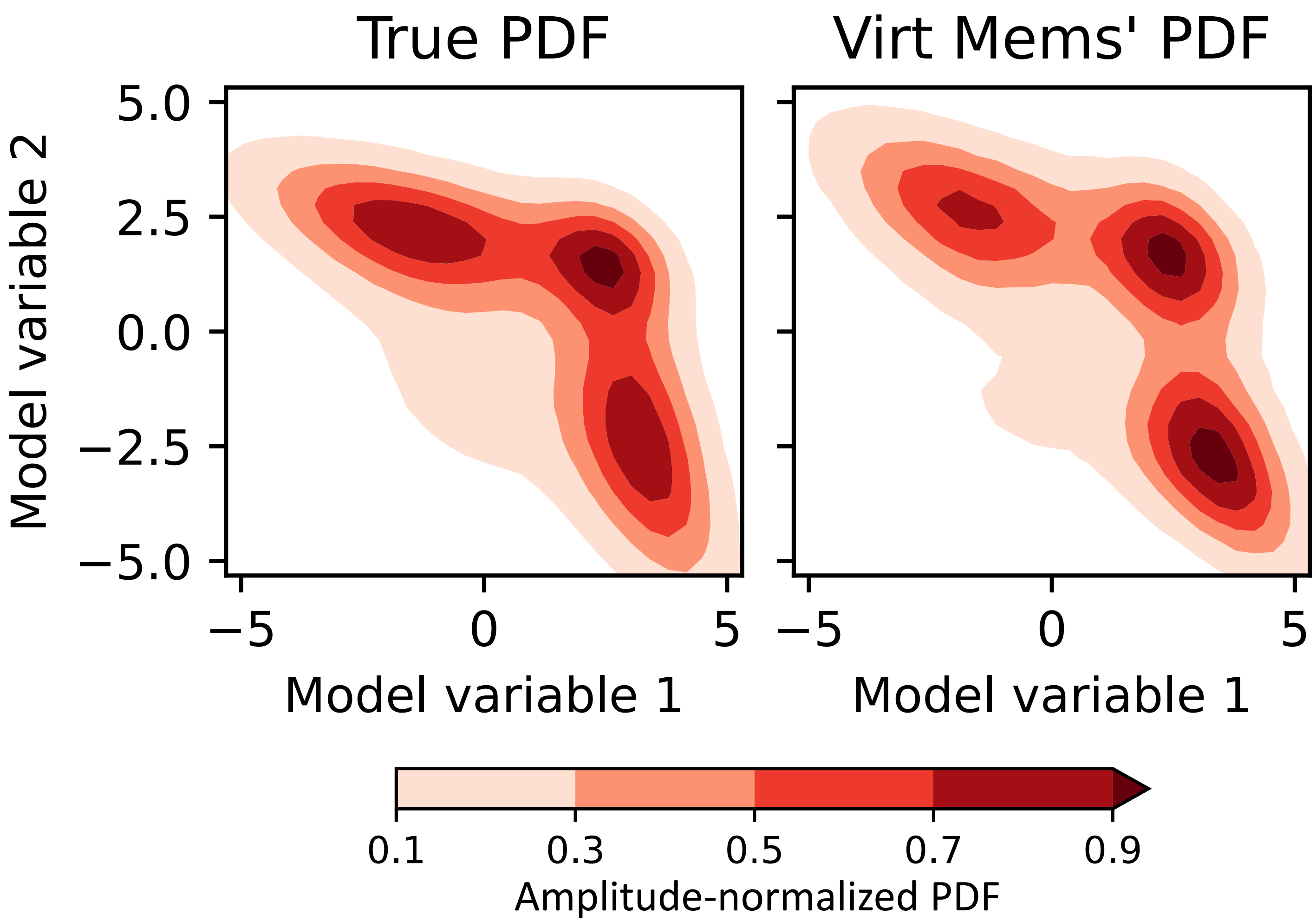


After PESE-GC



× Original members ● Virtual members

PESE-GC can handle non-Gaussian forecast distributions



Main Message

PESE-GC employs users' knowledge of 1D forecast PDFs* to create additional model-space ensemble members.

This reduces sampling errors, thus improving EnsDA**.

*   Aka, marginal forecast PDFs

** Tested using Lorenz 1996 model.

Main Message

PESE-GC employs users' knowledge of 1D forecast PDFs* to create additional model-space ensemble members.

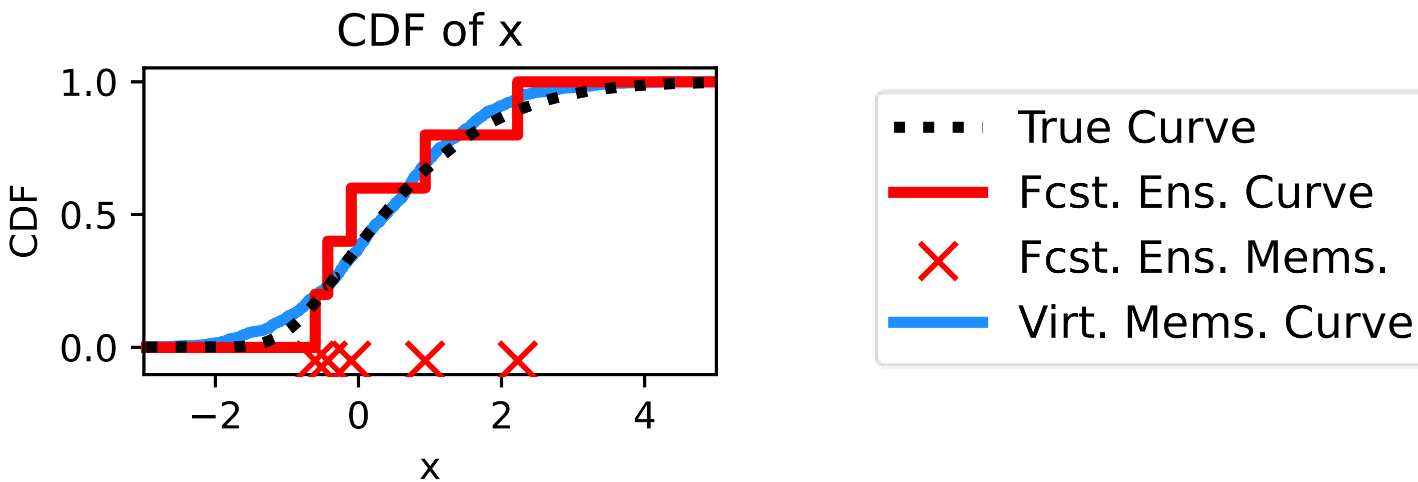
This **reduces sampling errors**, thus improving EnsDA**.

* Aka, marginal forecast PDFs

** Tested using Lorenz 1996 model.

PESE-GC reduces sampling errors

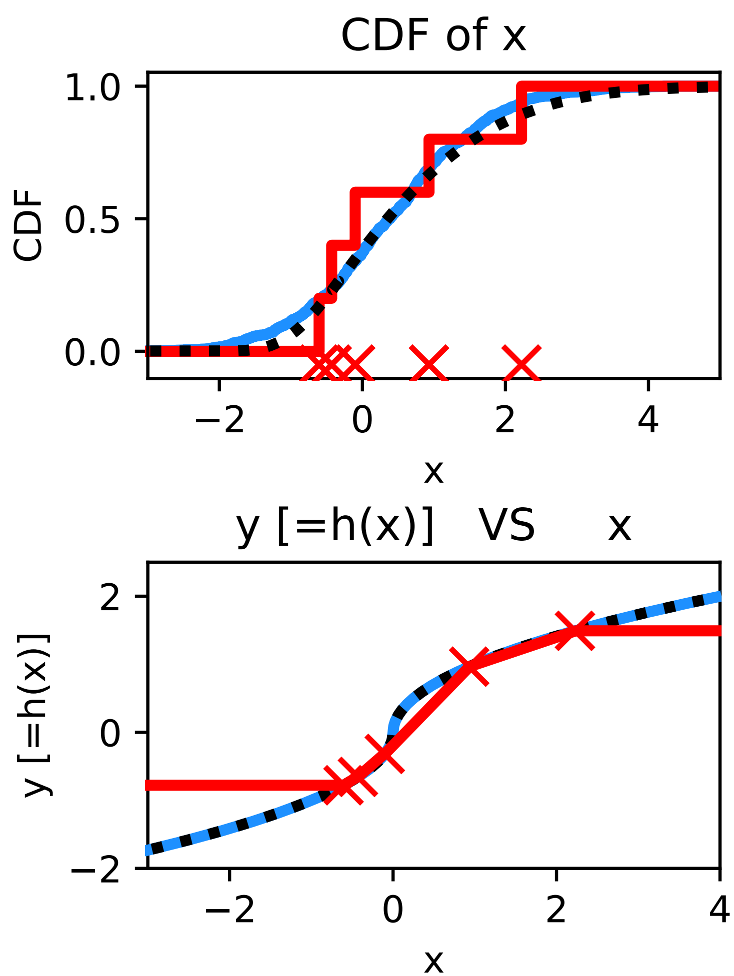
Bi-variate demonstration (1 model variable x, 1 obs variable y)

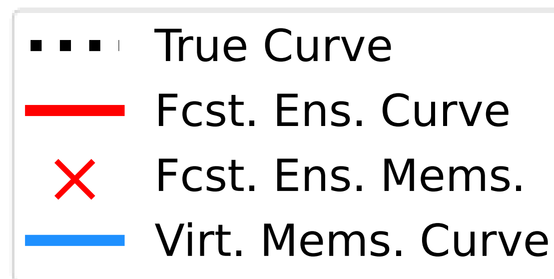


<u>Setup of demo:</u>

- True model-space forecast PDF is skewed normal
- 5 forecast members
- User knows forecast PDF is close to Gaussian, so **used PESE-GC with <u>Gaussian marginals</u>.**
- Obs operator: $h(x) = sign(x)\sqrt{|x|}$

PESE-GC reduces sampling errors in
1) Sampled relationship btwn obs & model quantities.



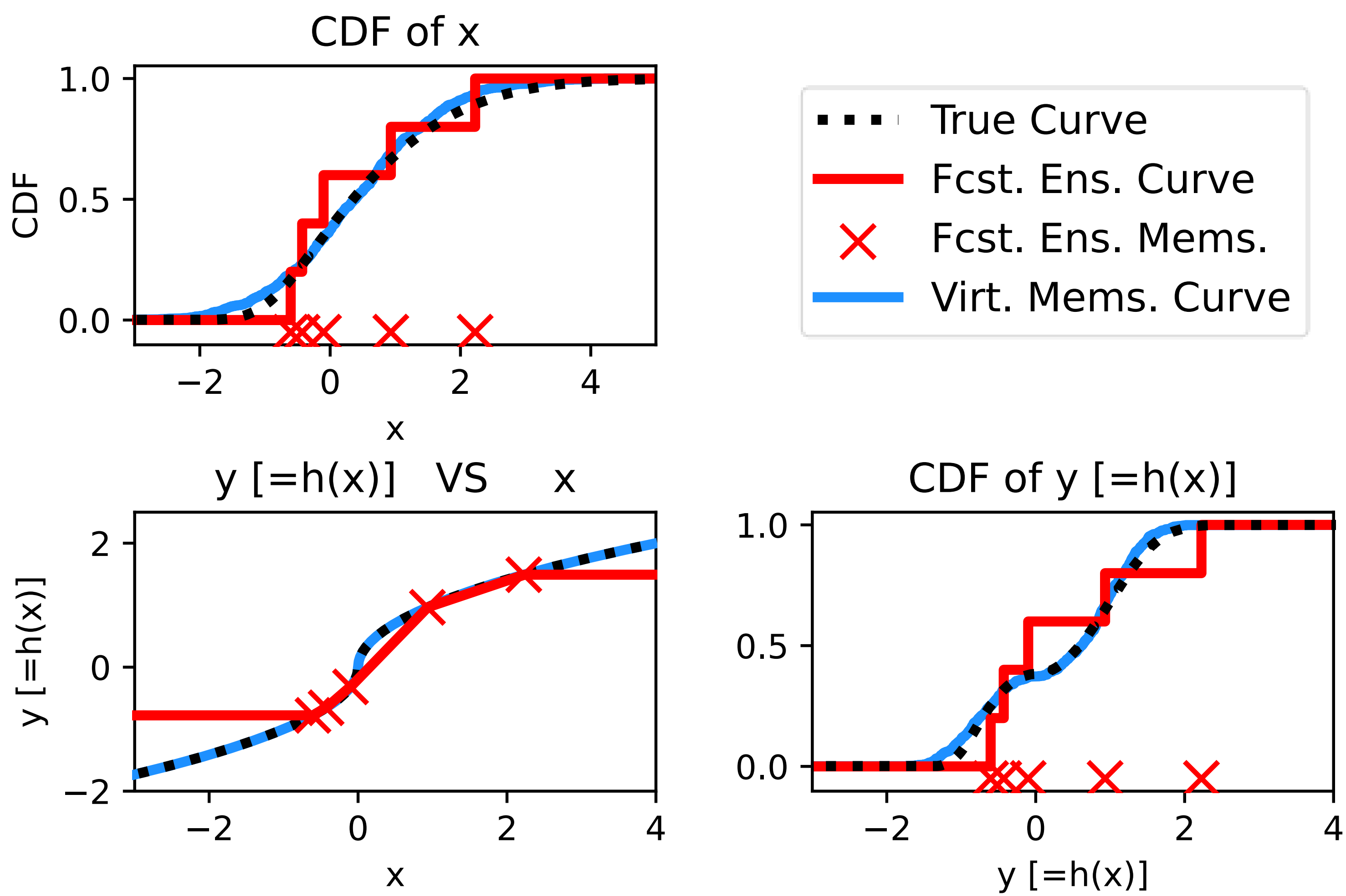


Virtual members' x-y curve is better than the forecast members' x-y curve.

Implies PESE-GC improves regression used by EnsDA.

PESE-GC reduces sampling errors in

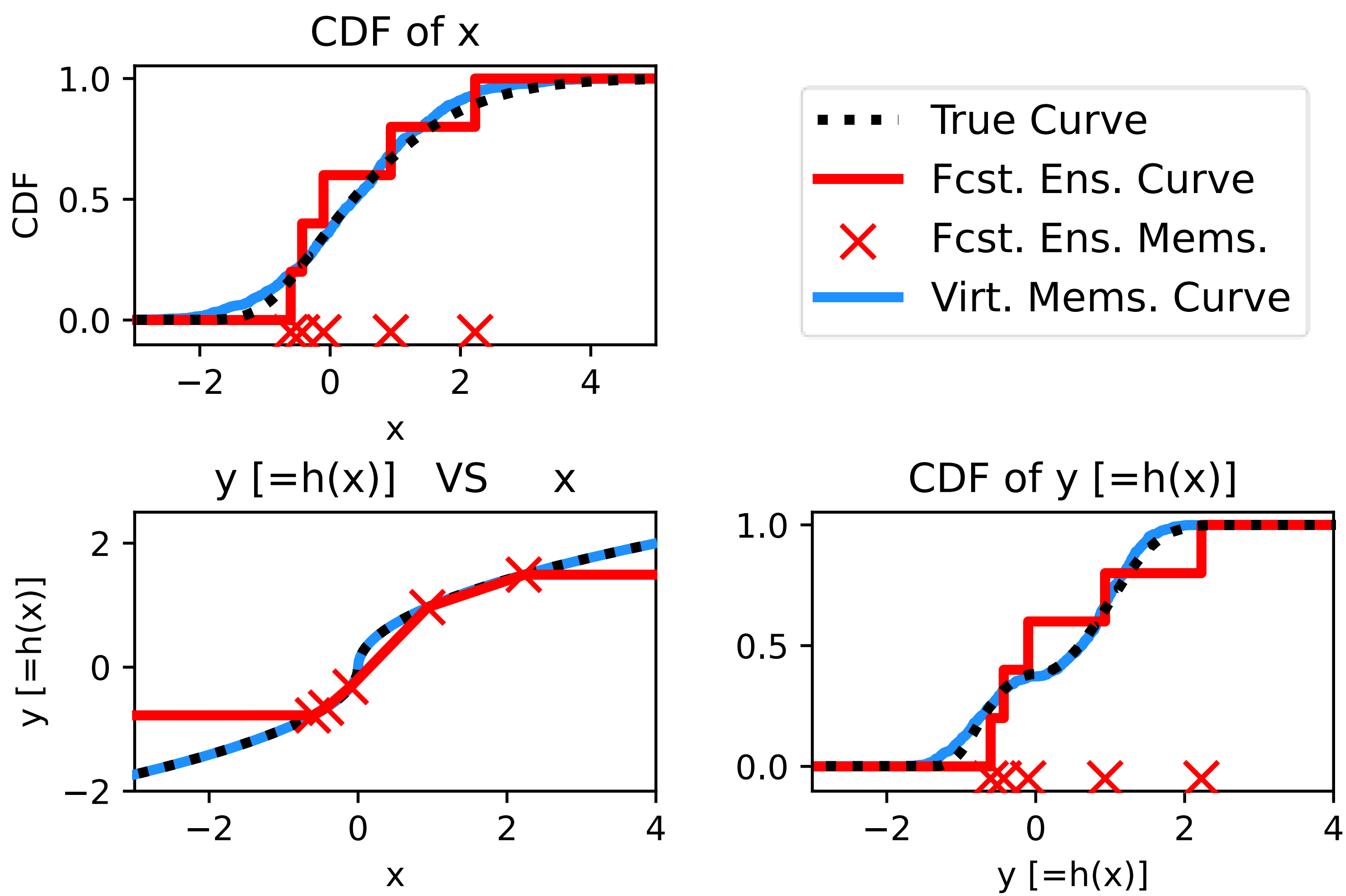
1) Sampled relationship btwn obs & model quantities.
2) Obs-space forecast statistics



PESE-GC reduces sampling errors in

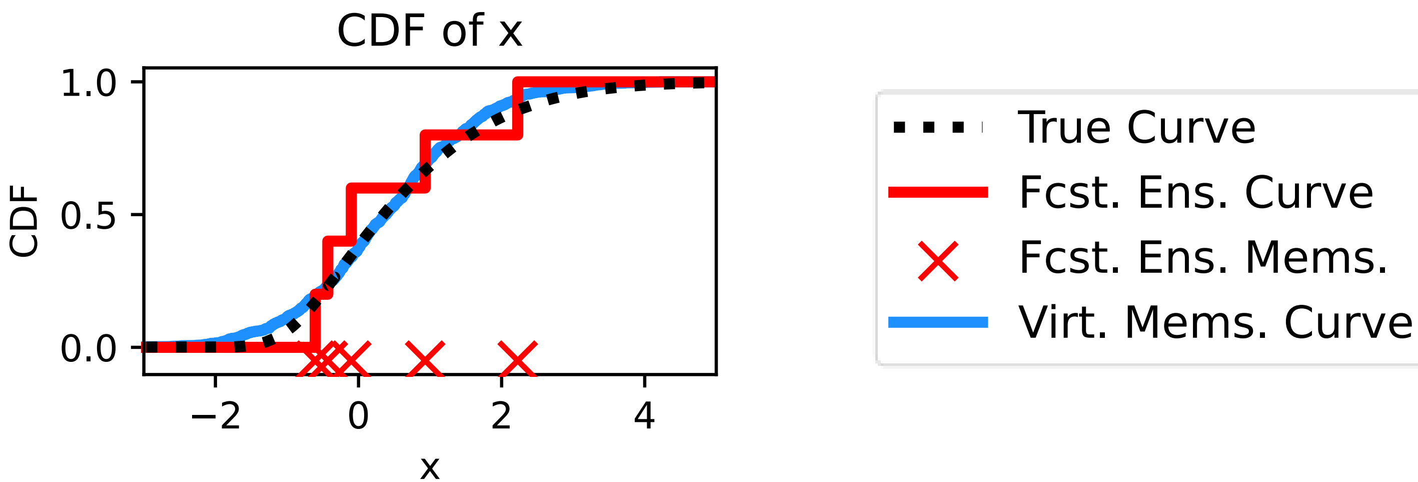
1) Sampled relationship btwn obs & model quantities.

2) Obs-space forecast statistics

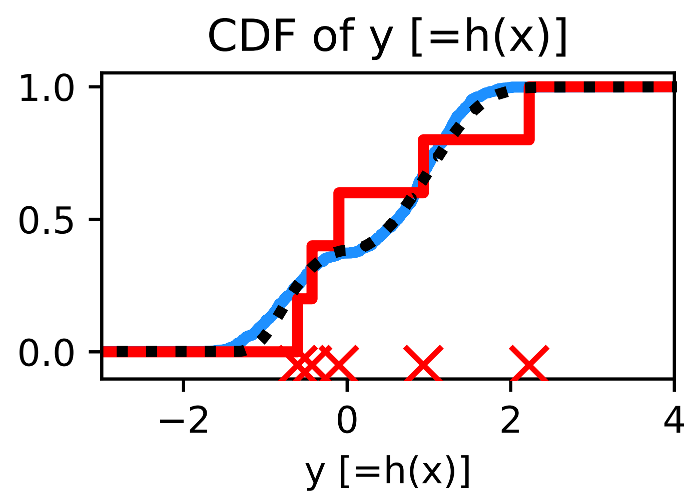


PESE-GC reduces sampling errors in

1) Sampled relationship btwn obs & model quantities.
2) Obs-space forecast statistics



Improving obs-space forecast statistics improves EnsDA.



Main Message

PESE-GC employs users' knowledge of 1D forecast PDFs* to create additional model-space ensemble members.

This **reduces sampling errors**, thus improving EnsDA**.

* Aka, marginal forecast PDFs

** Tested using Lorenz 1996 model.

Main Message

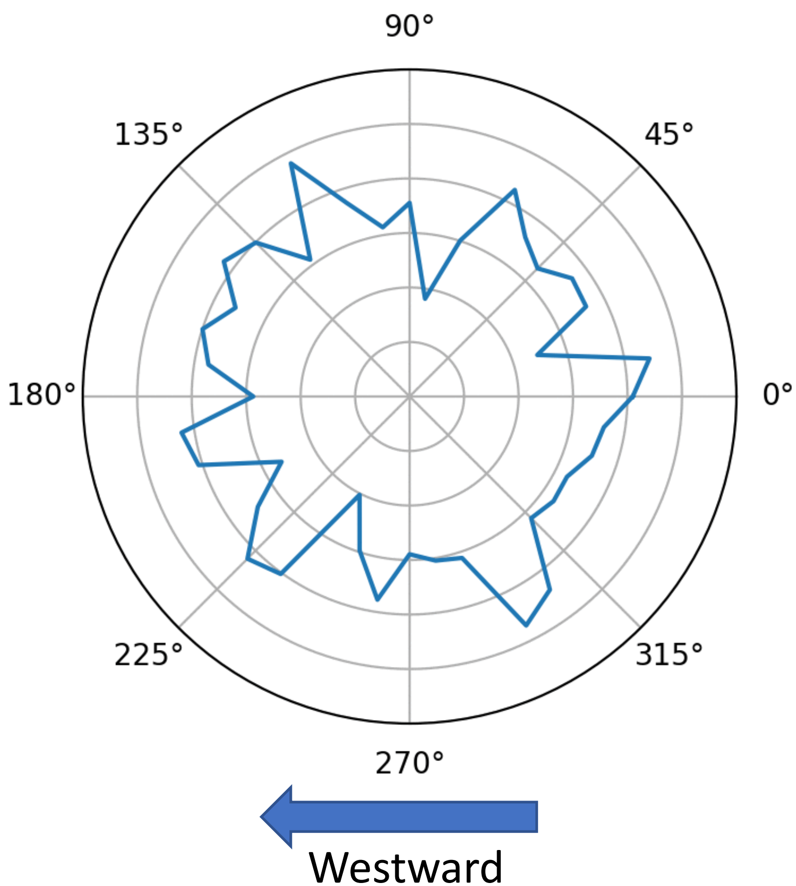
PESE-GC employs users' knowledge of 1D forecast PDFs* to create additional model-space ensemble members.

This reduces sampling errors, **thus improving EnsDA****.

*   Aka, marginal forecast PDFs

** Tested using Lorenz 1996 model.

PESE-GC is tested with 40-variable Lorenz 1996 “wave-on-a-ring” model



Setup of tests

Model settings: F=8.0, dt=0.05

Used NCAR’s Data Assimilation Research Testbed (DART)

EAKF is used as EnsDA algo*

PESE-GC assumes all marginal forecast PDFs are Gaussian

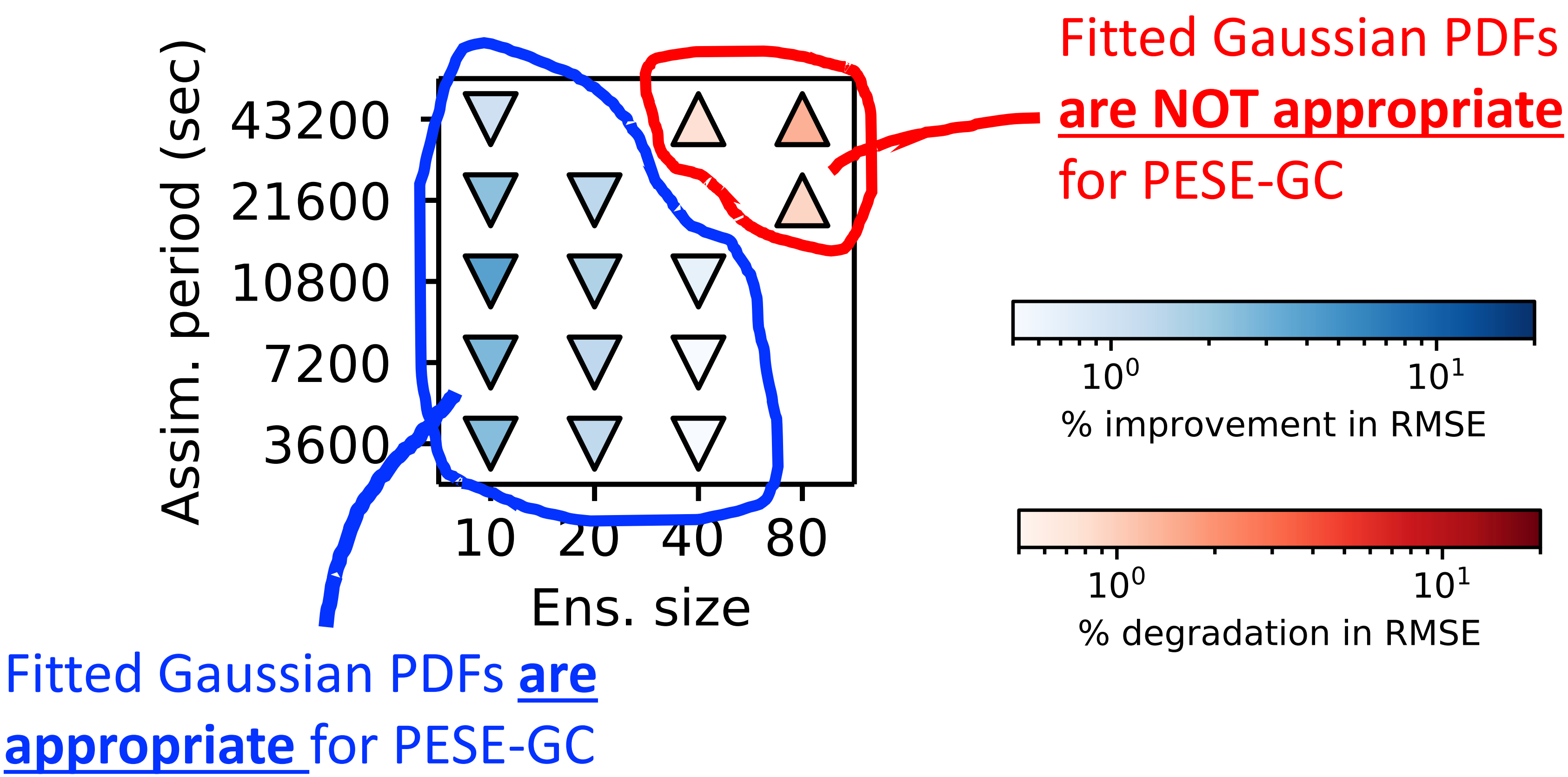
Obs op: $h(x) = sign(x)\sqrt{|x|}$

*Also tested with Perturbed Obs EnKF and Rank Histogram Filter with nonlinear regression

PESE-GC* improves EAKF when assumed marginal PDFs** are appropriate

*   PESE-GC increased ensemble size by a factor of 20

** Assumed PDFs are Gaussian in these tests.



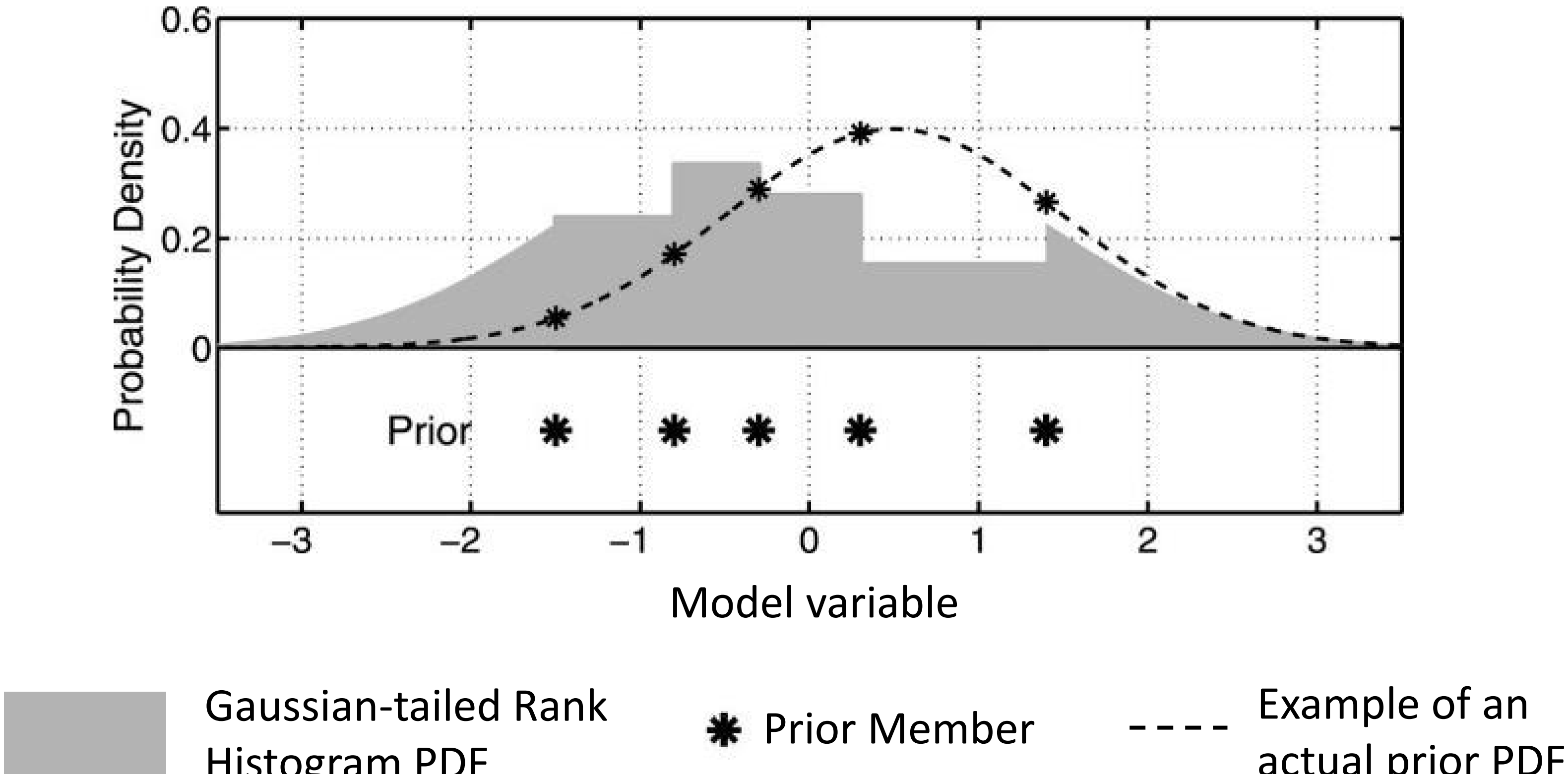
What if the user knows very little about the forecast marginals?

We can use non-parametric (i.e., "data-driven") approximations!

A non-parametric marginal PDF

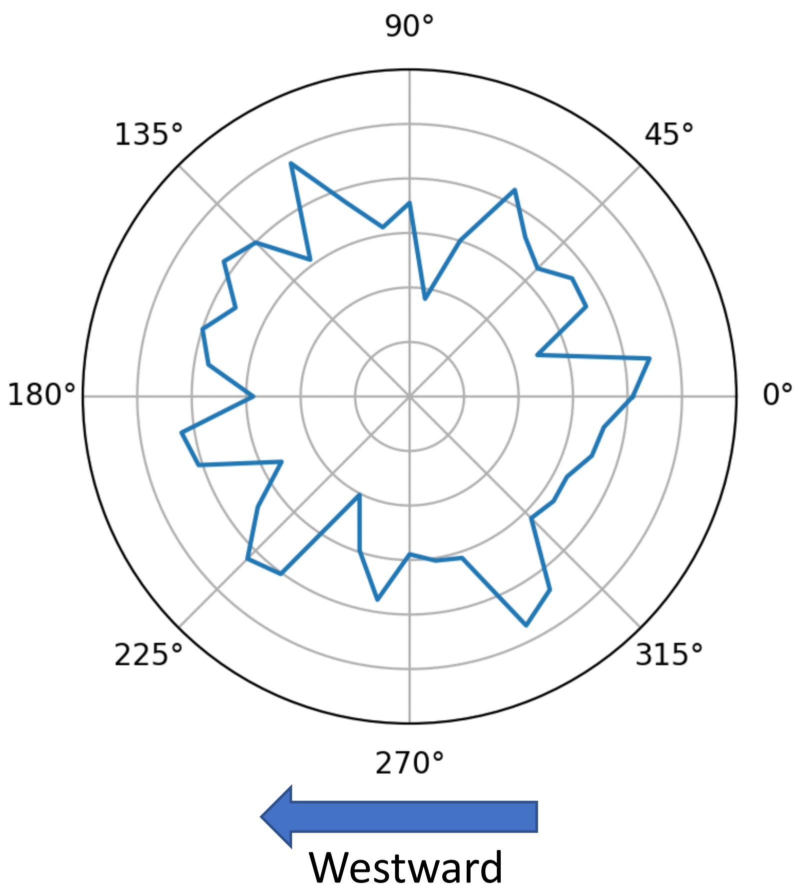
Gaussian-tailed Rank Histogram PDF

Visualization of Gaussian-tailed Rank Histogram PDF[1]



[1]Figure from Anderson, J. L., 2010: A Non-Gaussian Ensemble Filter Update for Data Assimilation. Mon. Wea. Rev., 138, 4186–4198, https://doi.org/10.1175/2010MWR3253.1.

PESE-GC is tested with 40-variable Lorenz 1996 “wave-on-a-ring” model



Setup of tests

Used NCAR’s Data Assimilation Research Testbed (DART)

Rank histogram filter (RHF) with nonlinear regression used as EnsDA algo*

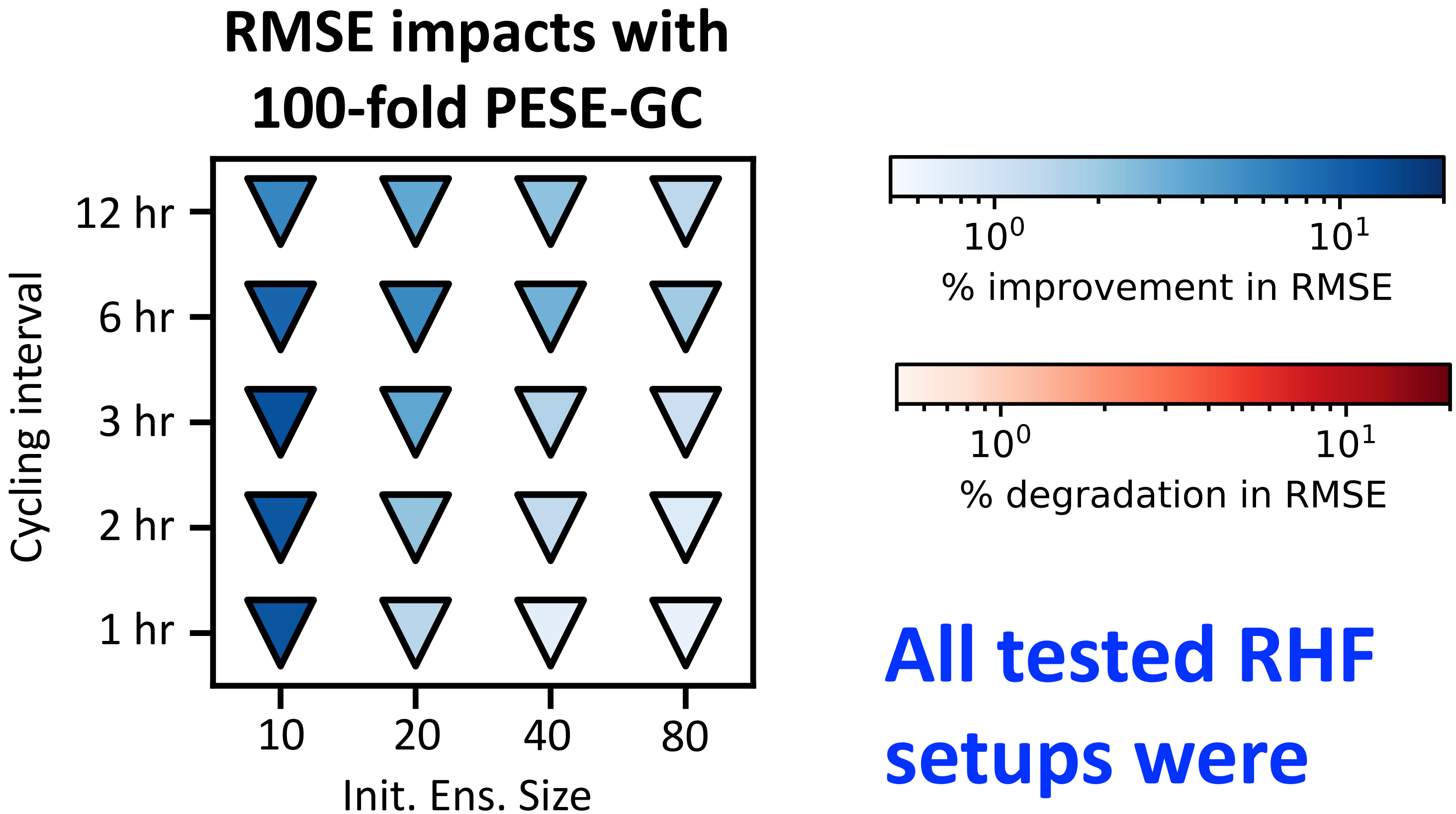
Non-parametric marginal PDFs used with PESE-GC

Obs op: $h(x) = sign(x)\sqrt{|x|}$

*Also tested with Perturbed Obs EnKF and EAKF.

PESE-GC improves RHF

These tests use PESE-GC with non-parametric marginals



All tested RHF setups were improved!

Main Message

PESE-GC employs users' knowledge of 1D forecast PDFs* to create additional model-space ensemble members.

This reduces sampling errors, **thus improving EnsDA****.

*   Aka, marginal forecast PDFs

** Tested using Lorenz 1996 model.

Main Message

PESE-GC employs users' knowledge of 1D forecast PDFs* to create additional model-space ensemble members.

This reduces sampling errors, thus improving EnsDA**.

* Aka, marginal forecast PDFs

** Tested using Lorenz 1996 model.

Avenues for future work

1. Build localization into PESE-GC [e.g., via ensemble modulation (Bishop and Hodyss, 2009)]

2. Test with realistic geophysical models

3. Does PESE-GC improve particle filter performance?

4. Can PESE-GC improve ML/AI in low-data situations?

Fin.

Thank you for your attention!

Happy to take questions & comments.

Main Message

PESE-GC employs users' knowledge of 1D forecast PDFs* to create additional model-space ensemble members.

This reduces sampling errors, thus improving EnsDA**.

* Aka, marginal forecast PDFs

** Tested using Lorenz 1996 model.

Manuscript submitted to Nonlinear Processes in Geophysics.