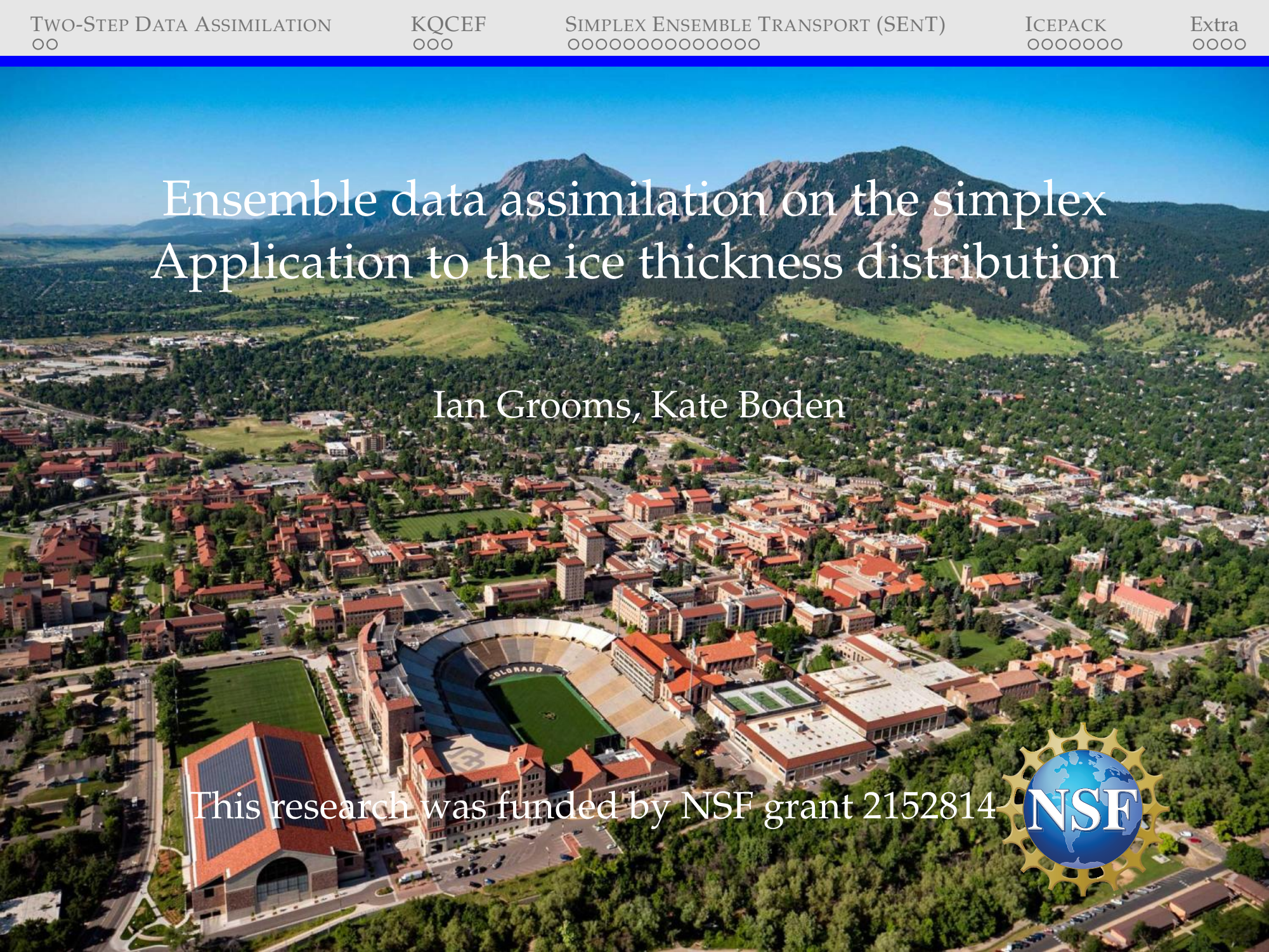
Ensemble data assimilation on the simplex
Application to the ice thickness distribution

Ian Grooms, Kate Boden



This research was funded by NSF grant 2152814

TWO-STEP ENSEMBLE FILTERS: REVIEW

Assume for the sake of exposition the standard observation model

$$\boldsymbol{y} = \mathcal{H}(\boldsymbol{x}) + \boldsymbol{\epsilon}, \quad \mathcal{N}(\mathbf{0}, \mathbf{R})$$

and define

$$\boldsymbol{z} = \mathcal{H}(\boldsymbol{x}).$$

First update the ensemble in observation space

$$\left\{\boldsymbol{z}_n^f\right\}_{n=1}^N, \boldsymbol{y}, \mathbf{R} \rightarrow \{\boldsymbol{z}_n^a\}_{n=1}^N$$

This first step can be any ensemble filter – EnKF, PF, whatever.

TWO-STEP ENSEMBLE FILTERS: REVIEW

Then in update the ensemble in model/state space

$$\left\{\boldsymbol{x}_n^f\right\}_{n=1}^N, \left\{\boldsymbol{z}_n^f\right\}_{n=1}^N, \boldsymbol{z}_n^a \to \boldsymbol{x}_n^a.$$

The second step samples $\boldsymbol{x}_n^a$ from the conditional distribution $\boldsymbol{X}|\mathcal{H}(\boldsymbol{X}) = \boldsymbol{z}_n^a$.

The first two-step ensemble filter

Anderson, *A local least-squares framework for ensemble filtering*, MWR 2003.

The connection between two-step sampling algorithm and the Bayesian posterior

Grooms, *A comparison of nonlinear extensions to the ensemble Kalman filter*, Comp. Geo. 2022.

TWO-STEP ENSEMBLE FILTERS: REVIEW

Then in update the ensemble in model/state space

$$\left\{\boldsymbol{x}_n^f\right\}_{n=1}^N, \left\{\boldsymbol{z}_n^f\right\}_{n=1}^N, \boldsymbol{z}_n^a \to \boldsymbol{x}_n^a.$$

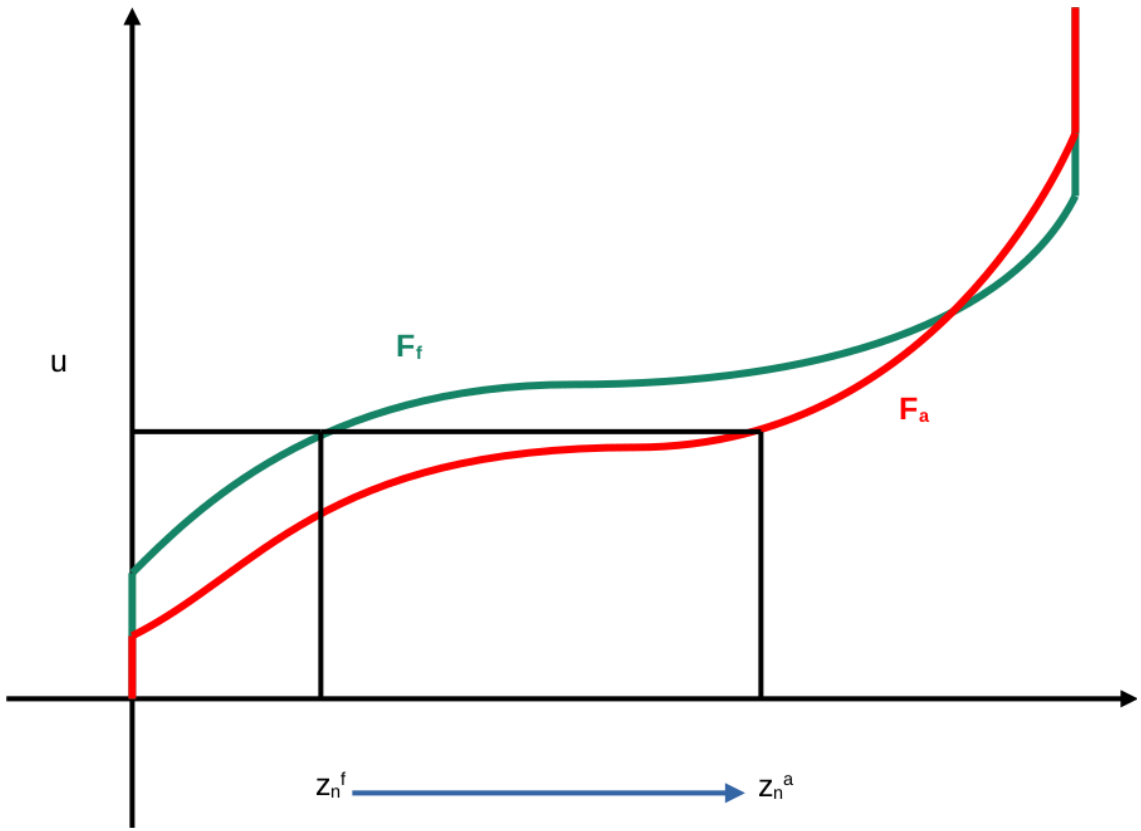
The second step samples $\boldsymbol{x}_n^a$ from the conditional distribution $\boldsymbol{X}|\mathcal{H}(\boldsymbol{X}) = \boldsymbol{z}_n^a$.

The first two-step ensemble filter
Anderson, *A local least-squares framework for ensemble filtering*, MWR 2003.

The connection between two-step sampling algorithm and the Bayesian posterior
Grooms, *A comparison of nonlinear extensions to the ensemble Kalman filter*, Comp. Geo. 2022.

Recall the Quantile-Conserving Ensemble Filter (QCEF), composing a probability integral transform and an inverse sampling transform

$$z_n^a = F_a^{-1}\left(F_f\left(z_n^f\right)\right).$$



$F_{a,f}$ are the analysis/forecast cdfs in observation space.

I developed a non-parametric QCEF modeling the prior pdf as a sum of

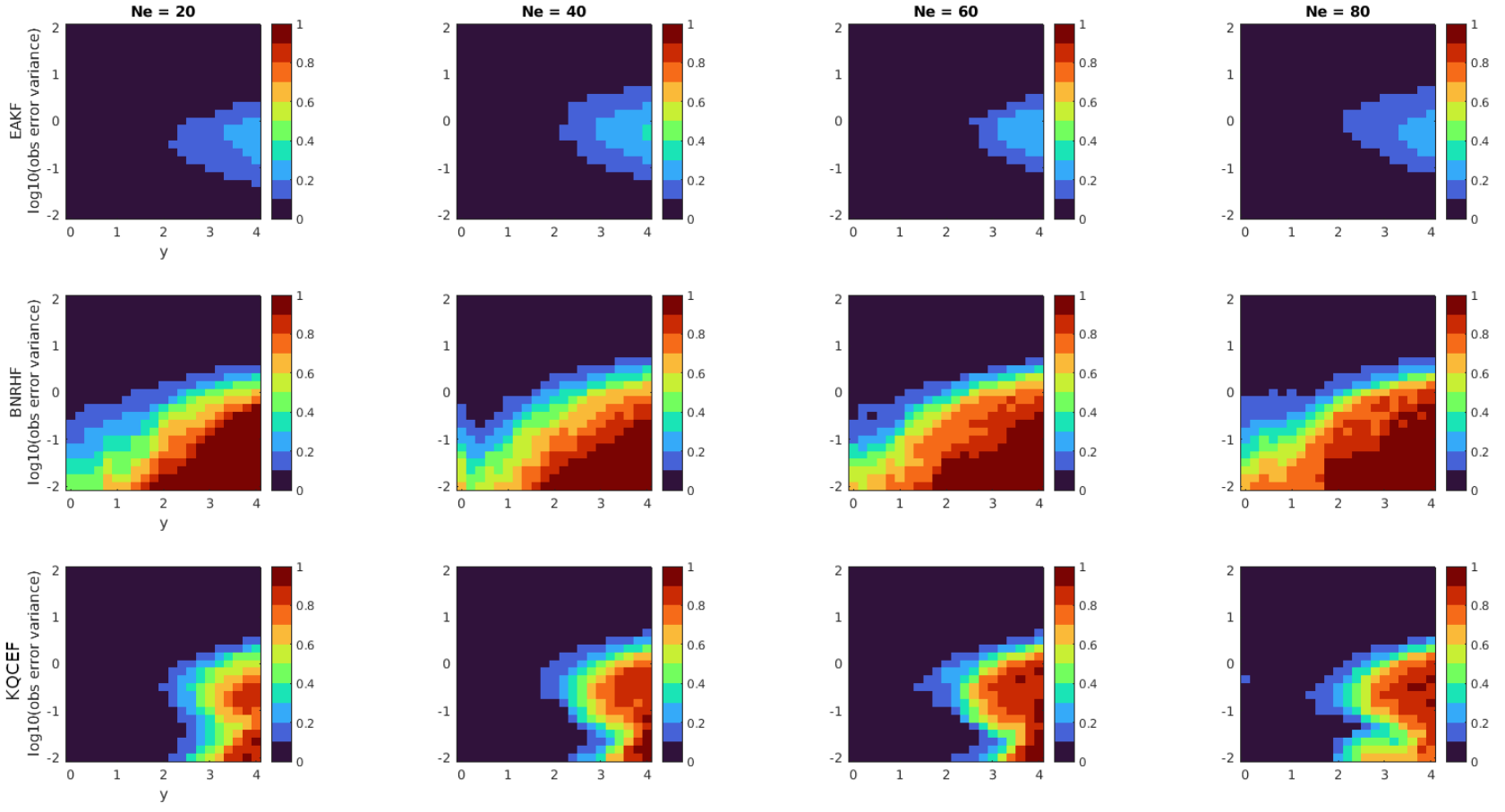
- Dirac delta distributions on the boundaries, plus
- A smooth pdf between the boundaries.

The weights on each component are estimated from the forecast ensemble, and the interior pdf is estimated using kernel density estimation (KDE) with boundary corrections.

I evaluate F_a^{-1} and F_f using a combination of quadrature and rootfinding.

Grooms & Riedel, *A Quantile-Conserving Ensemble Filter Based on Kernel-Density Estimation*, Remote Sensing 2024.

I have implemented this in DART, where it is now an option in the main release together with EAKF, RHF, etc.



Each panel shows the fraction of 100 experiments where the null hypothesis that the analysis ensemble was drawn from the known true posterior distribution was rejected by the Kolmogorov-Smirnov test at the 5% significance level. The prior is a standard normal. The likelihood is normal with mean y that varies along the horizontal axis and variance γ^2 that varies along the vertical axis.

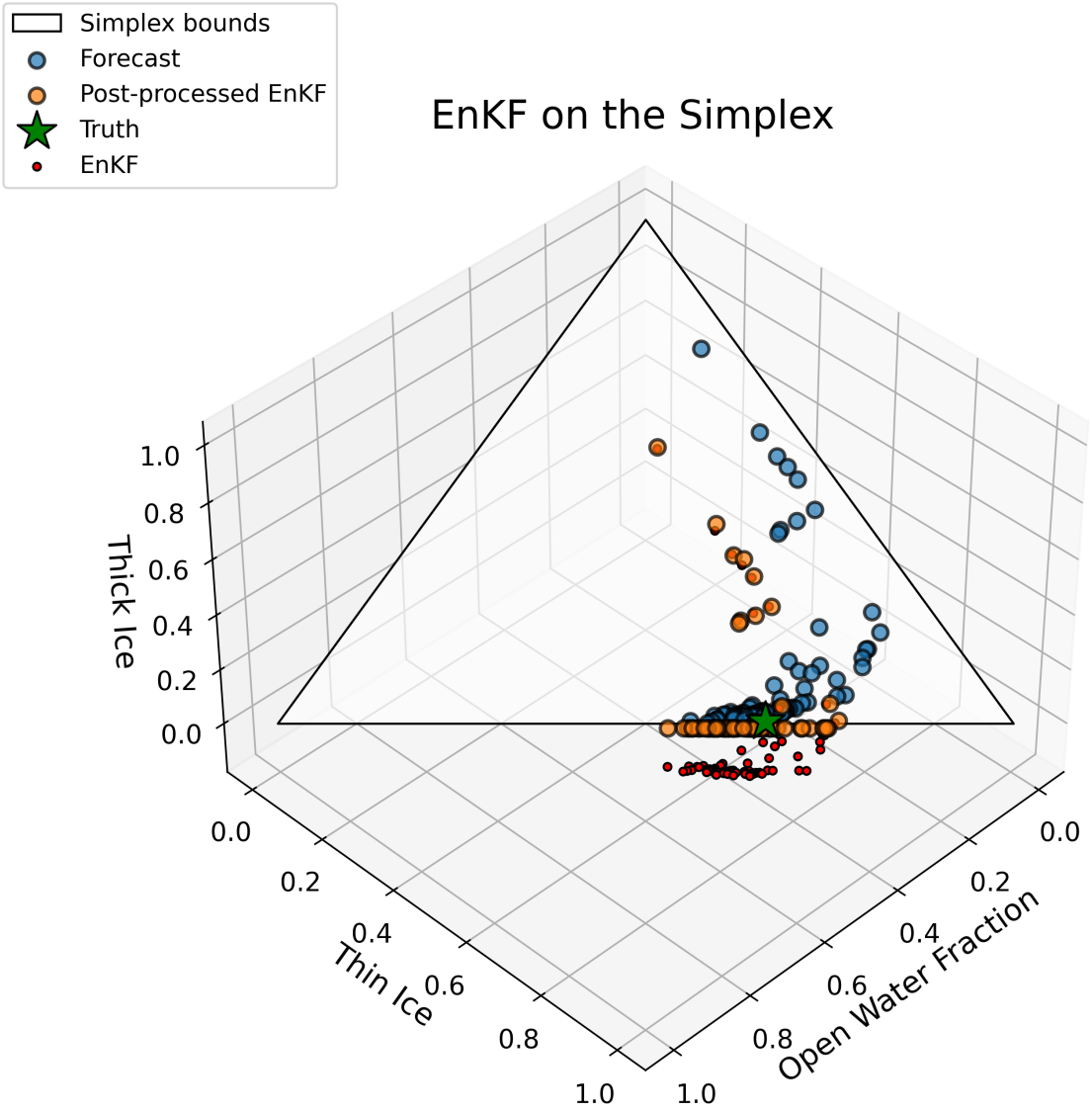
Many sea ice models have, on each grid cell, an Ice Thickness Distribution (ITD): The total area of a grid cell is divided among N_{cat} thickness categories, each having fractional area

$$a_i, \quad i = 1, \ldots, N_{\text{cat}}.$$

The areas are constrained to live on the simplex

$$a_i \geq 0 \text{ and } \sum_{i=0}^{N_{\text{cat}}} a_i = 1.$$

Simplex constraints appear in other contexts as well, including epidemiological, ecological, and biogeochemical models.



Example using $N_{\text{cat}} = 2$.

We observe Sea Ice Concentration (SIC)

$$\text{SIC} = 1 - a_0$$

where a_0 is the fractional area of a grid cell covered by open water. We use KQCEF.

In the second step we must sample from the conditional distribution of a_i, $i = 1, \ldots, N_{\text{cat}}$ given a_0.

We model the joint distribution of $a_0, \ldots, a_{N_{\text{cat}}}$ using Dirichlet distributions, which are natively supported on the simplex.

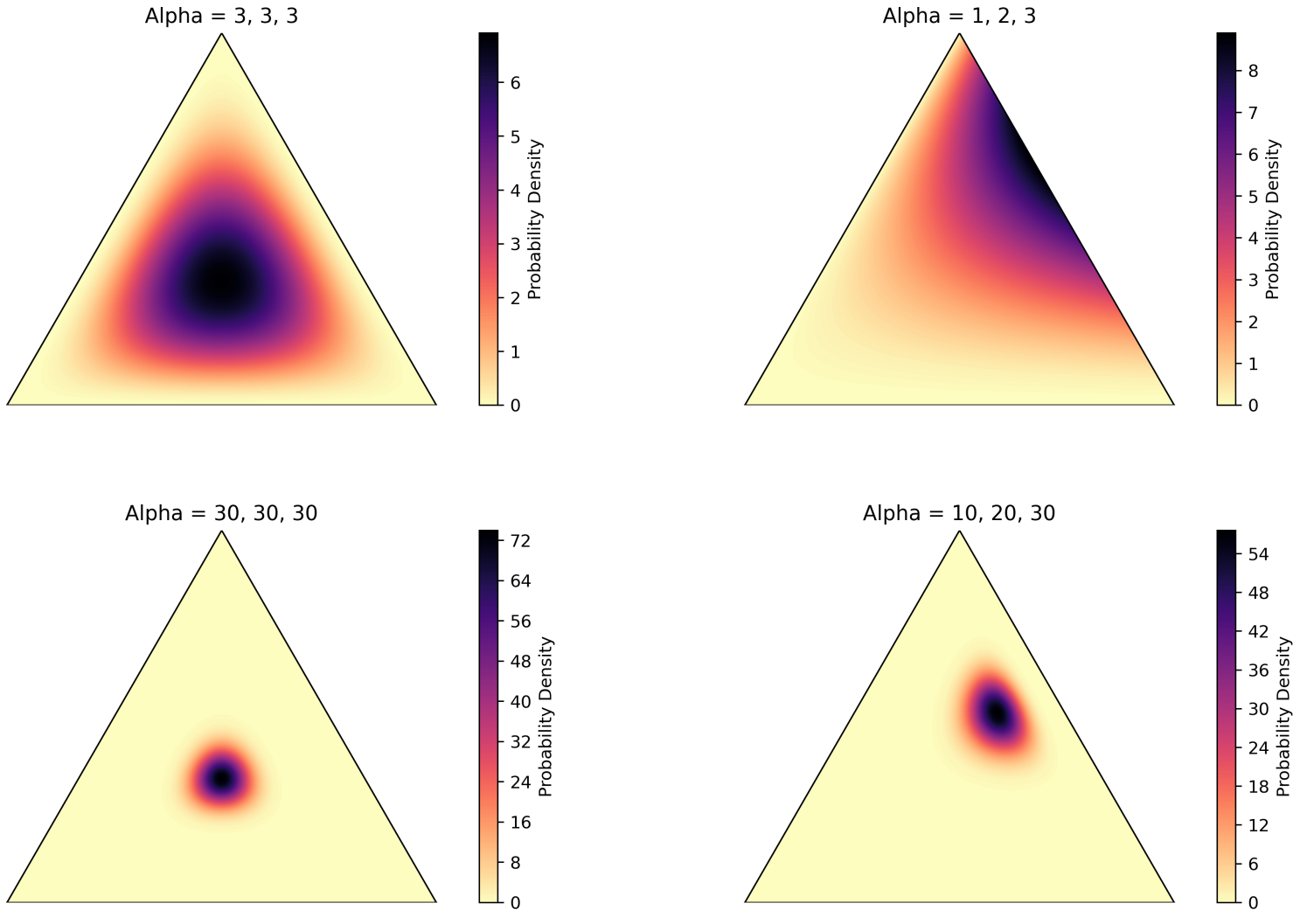
We observe Sea Ice Concentration (SIC)

$$\text{SIC} = 1 - a_0$$

where a_0 is the fractional area of a grid cell covered by open water. We use KQCEF.

In the second step we must sample from the conditional distribution of a_i, $i = 1, \ldots, N_{\text{cat}}$ given a_0.

We model the joint distribution of $a_0, \ldots, a_{N_{\text{cat}}}$ using Dirichlet distributions, which are natively supported on the simplex.



Examples of Dirichlet distributions.

We model the joint distribution of $a_0, \ldots, a_{N_{\text{cat}}}$ directly as

$$p_{a_0,\boldsymbol{a}}(a_0, \ldots, a_{N_{\text{cat}}}) = \sum_{k=1}^{K} \pi_k \text{Dirichlet}(a_0, \boldsymbol{a}; \alpha_0, \boldsymbol{\alpha}_k)$$

where there are $K = 2^{N_{\text{cat}}+1} - 1$ *classes*; one on the simplex and one on each element of the boundary of the simplex (which is itself a lower-dimensional simplex).

Each class is a Dirichlet distribution with parameters $\alpha_0, \boldsymbol{\alpha}_k$.

We could in principle have more than one Dirichlet per simplex, but have not pursued that yet.

The MLE estimate for the weights is

$$\pi_k \approx \frac{N_k}{N_e}$$

where N_k is the number of ensemble members in class k. In practice the number of classes with nonzero weights is manageable.

$\boldsymbol{\alpha}_k$ and $\alpha_{0,k}$ are estimated by iteratively maximizing the likelihood in class k.

For a mixture of Dirichlets, the conditional is also a mixture of Dirichlets where the parameters $(\pi_k, \boldsymbol{\alpha}_k)$ depend in a known, explicit way on the joint parameters:

$$p_{\boldsymbol{a}|a_0}(\boldsymbol{a}) = \sum_{k=1}^{K} \pi_k(a_0) \text{Dirichlet}\left(\frac{\boldsymbol{a}}{1-a_0}; \boldsymbol{\alpha}_k\right).$$

As the notation suggests, conditioning does not change $\boldsymbol{\alpha}_k$, but it does change the weights.

To sample from this conditional distribution we could first randomly choose which class the member will be in, then randomly sample within that class; there are efficient algorithms for both these steps.

The problem with this is that we (eventually) want to do this on a grid, not just at one grid point. If we sampled randomly at each grid point, the analysis ensemble would be spatially incoherent.

To maintain spatial coherence we adopt a *transport* approach.

We will define a (random) map that takes in a forecast ensemble member and puts out an analysis ensemble member.

By using the same map at all grid points we maintain spatial coherence.

The map has two parts: a map from class to class, followed by a map from simplex to simplex.

If forecast ensemble member n is in class k_n^f, then we let the analysis class k_n^a be a draw from the k_n^f-th column of matrix $\mathbf{A}$ that solves

$$\text{minimize} \sum_{i,j} d_{ij} a_{ij}$$

$$\text{subject to } a_{ij} \geq 0,\ \sum_i a_{ij} = 1,\ \mathbf{A}\boldsymbol{\pi}^f = \boldsymbol{\pi}^a$$

where $\boldsymbol{\pi}^{f,a}$ are the forecast/analysis class weights and d_{ij} are distances between classes.

This requires the solution of a linear programming problem for each ensemble member, but the size of $\mathbf{A}$ is usually quite small (maybe 5×5 at most), which keeps costs down.

Once we have decided which class the n-th analysis ensemble member will come from, we transport the forecast value to the analysis value using

$$\frac{\boldsymbol{a}_n^a}{1 - a_{0,n}^a} = \boldsymbol{F}_a^{-1}\left(\boldsymbol{F}_f\left(\frac{\boldsymbol{a}_n^f}{1 - a_{0,n}^f}\right)\right).$$

Since we are in N_{cat} dimensions $\boldsymbol{F}_{f,a}$ are not cdfs, but the idea is the same as the 1D case.

Any joint pdf can be factored as a chain of conditionals and marginals

$$p_{\boldsymbol{x}}(x_1, \ldots, x_n) = p_{x_1}(x_1) p_{x_2|x_1}(x_2) \cdots p_{x_n|x_1,\ldots,x_{n-1}}(x_n).$$

Let F_1 be the cdf of p_{x_1}, and F_i be the cdf of $p_{x_i|x_{j<i}}$. Define

$$\boldsymbol{F} = \begin{pmatrix} F_1 \\ \vdots \\ F_n \end{pmatrix}.$$

Then $\boldsymbol{F}(\boldsymbol{X}) = \boldsymbol{U}$ is uniformly distributed on $[0,1]^n$.

Conversely, if $\boldsymbol{U}$ is uniform on $[0,1]^n$ then $\boldsymbol{X} = \boldsymbol{F}^{-1}(\boldsymbol{U})$ has pdf $p_{\boldsymbol{x}}(x_1, \ldots, x_n)$.

If we compose the forward and inverse transforms we can transport realizations from one distribution to another.

These are triangular transport maps.

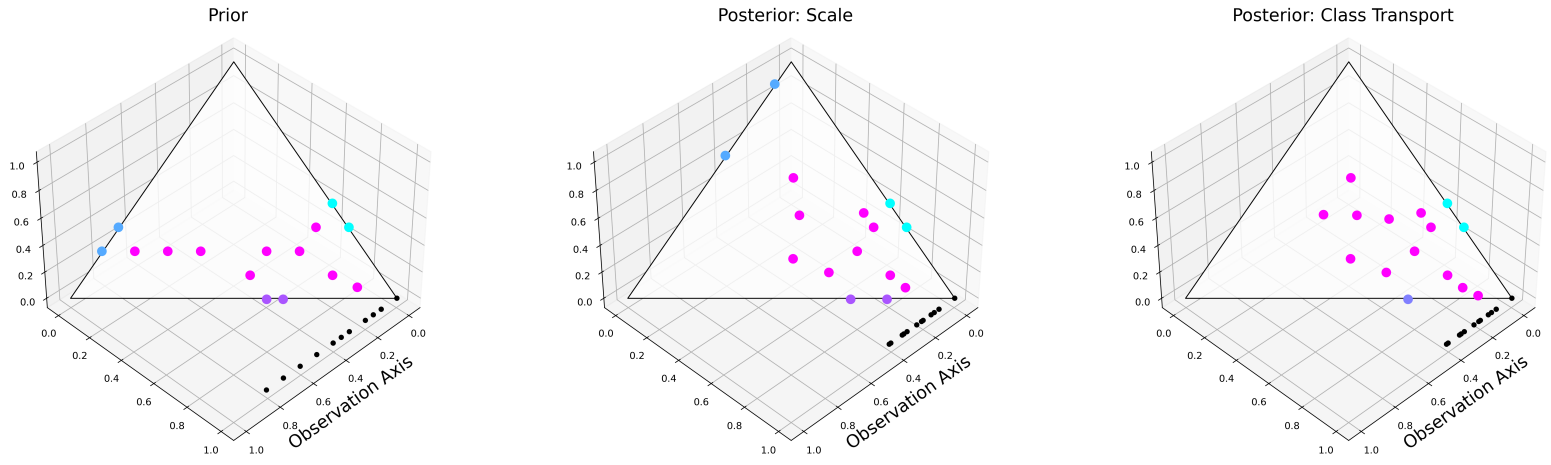
Thanks to our parametric assumption, every scalar distribution $X_i|X_{j<i}$ is a Beta distribution with known parameters.

If the ensemble member stays in the same class, then this transport can be shown to reduce to scaling:

$$\boldsymbol{a}^a = \frac{1 - a_0^a}{1 - a_0^f} \boldsymbol{a}^f$$

which has been used previously for sea ice.

(If the ensemble member does change class then we are mapping between different simplexes. We do this by always mapping first to $[0, 1]^{N_{\text{cat}}}$, filling in with random draws from a uniform when necessary, and then mapping from $[0, 1]^{N_{\text{cat}}}$ to the target simplex, ignoring unused values of $\boldsymbol{u}$.)

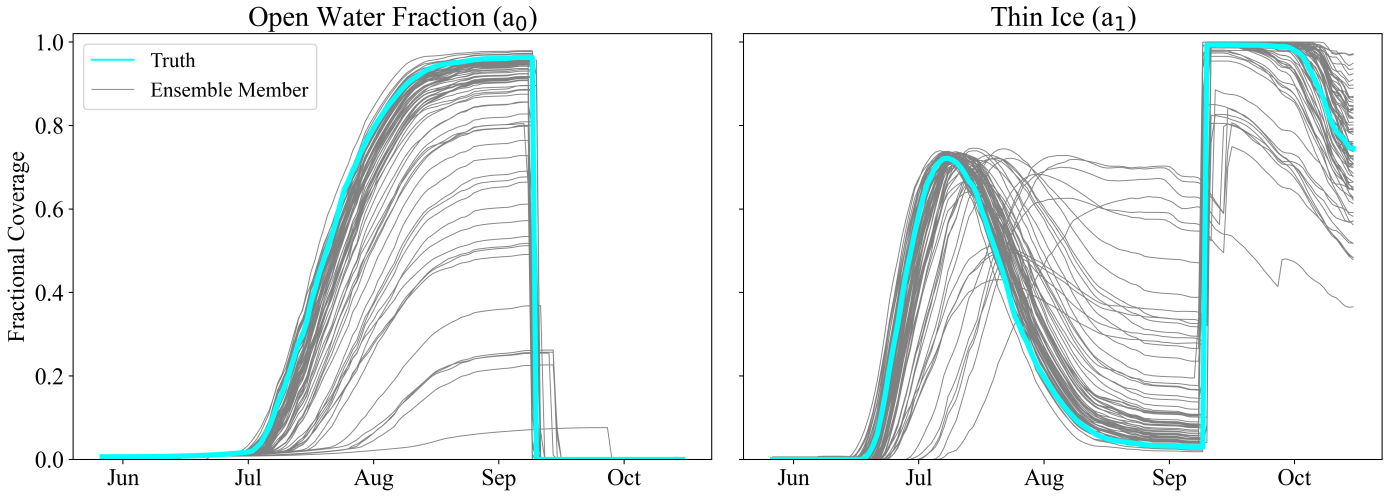


Colors indicate classes; black dots indicate observation space values. Left: Prior. Center: Scaling. Right: Class Transport.

We test this in the 'single-column' version of CICE5, called Icepack. We use $N_{\text{cat}} = 5$ thickness categories (default).

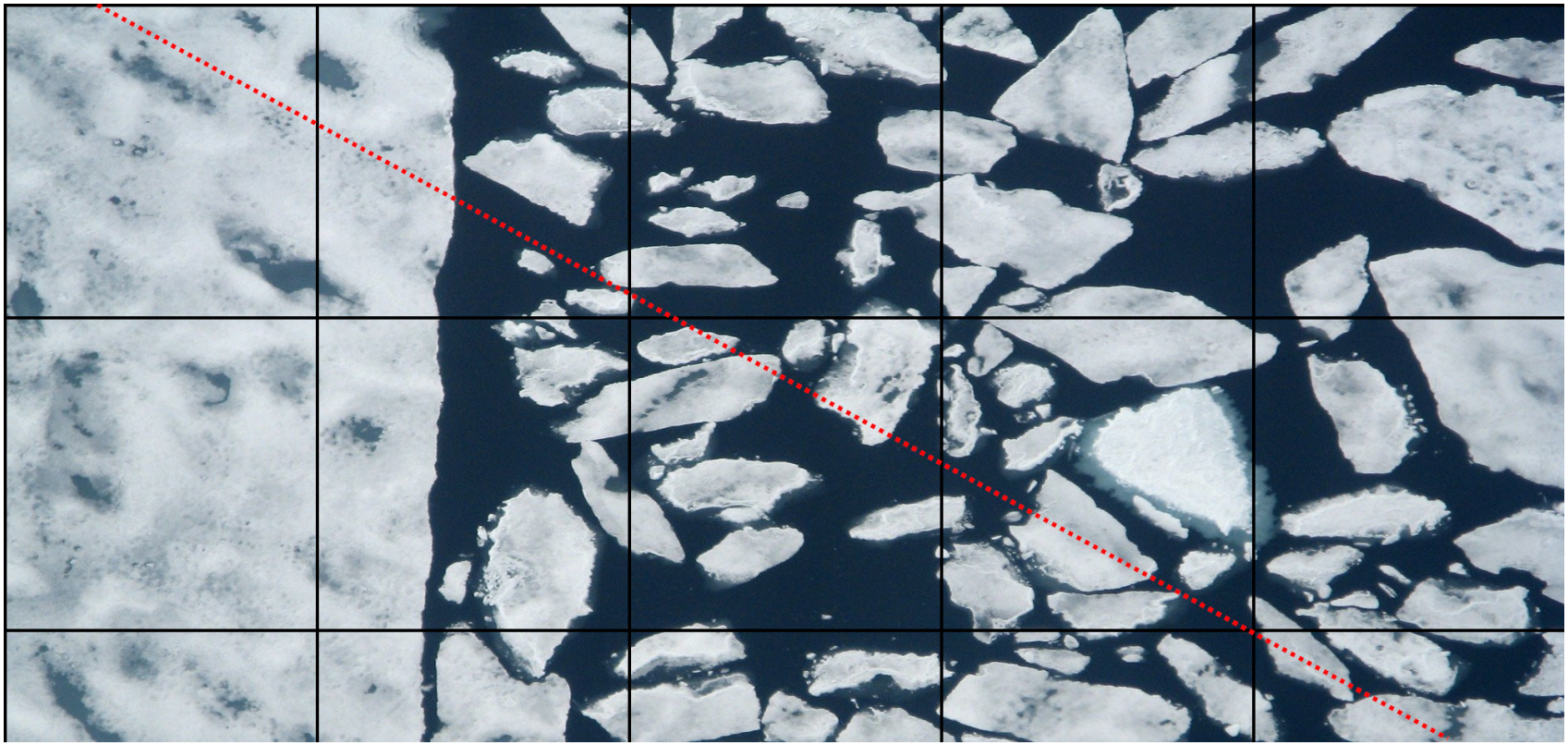
This model has no intrinsic variability, so we force using the 80-member DART+CAM6 reanalysis.

That's still not enough variability, so we further perturb parameters related to snow grain radius and thermal conductivity.



Results of a free run (no assimilation). The reference/truth case is marked in light blue.

What is the likelihood function for SIC observations?



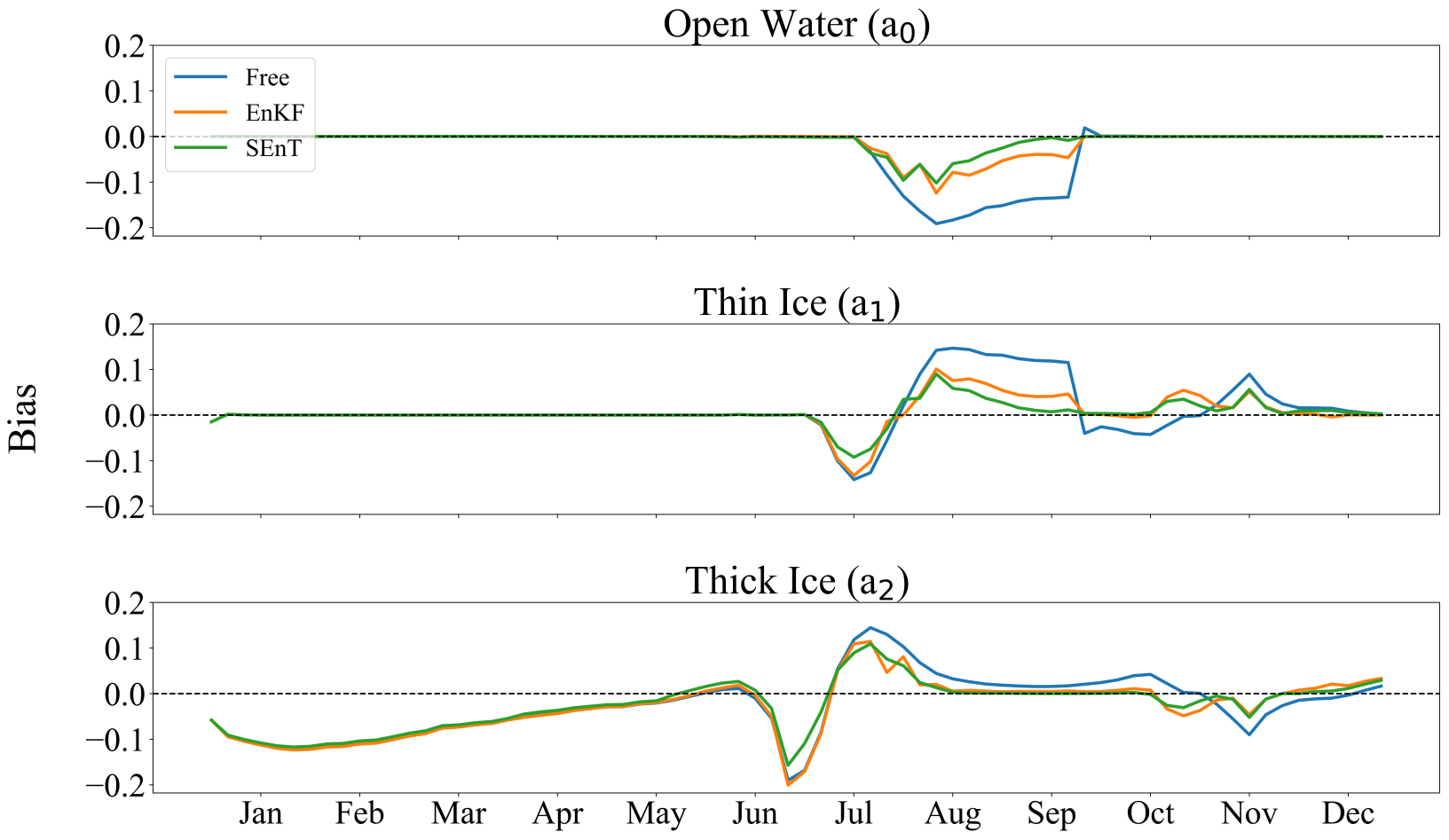
For satellites like CryoSat and ICESat-2 we can model the observations on a single grid cell as N_o independent draws, k of which see ice and $N_o - k$ of which see open water. The observation distribution is therefore Binomial, and the likelihood is Beta.

We assimilate observations with $N_o = 10$, which leaves a fairly large uncertainty.

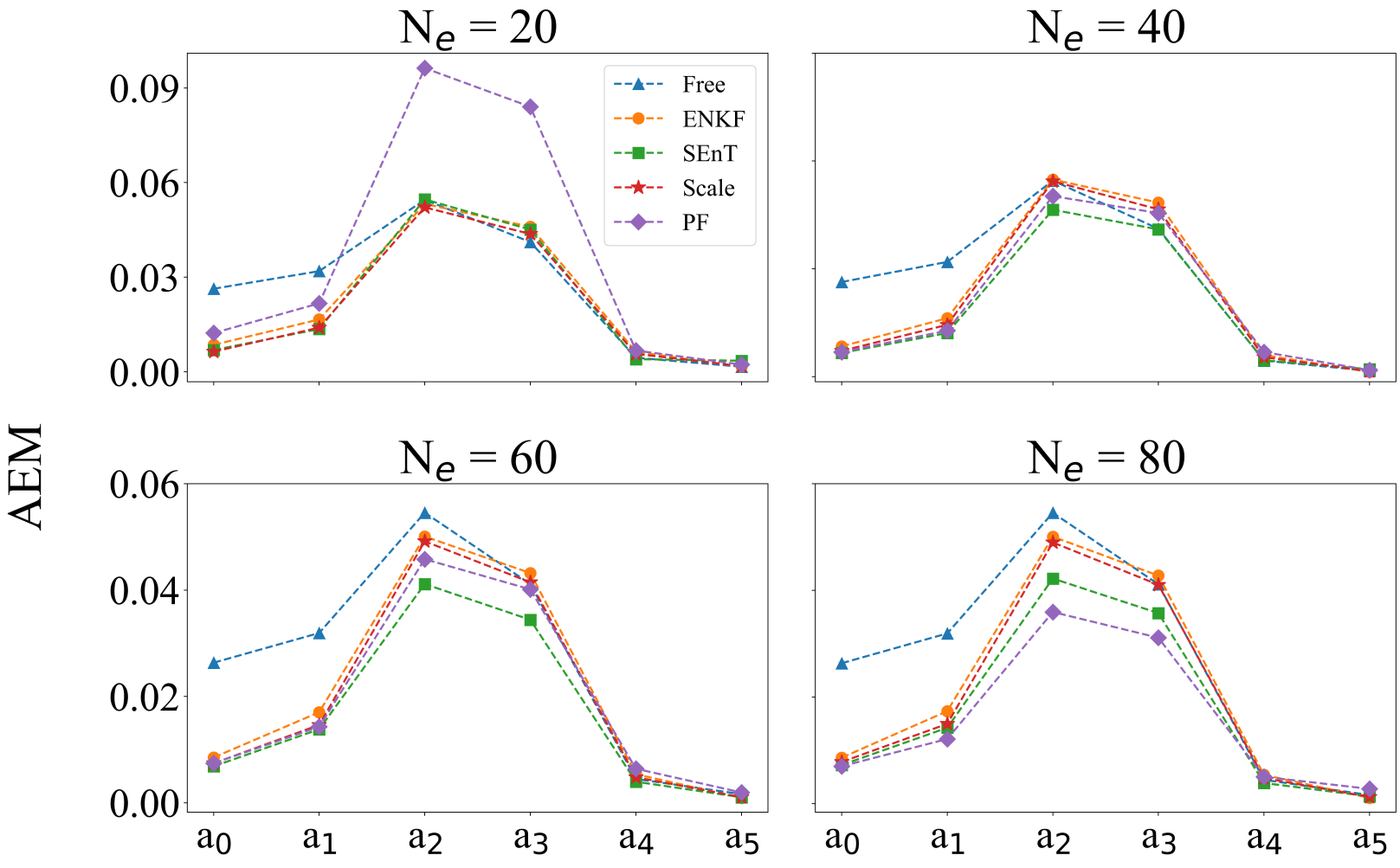
If you assimilate daily the spread drops quickly and you don't see much difference between methods.

We assimilate every 5 days, but for one year that only gives us 73 cycles. So we perform 5 experiments, each with a shifted 5-day cycle, so that we cover all days of the year.

We compare against an ESRF with postprocessing, a standard particle filter (PF), and a two-step method where the second step just scales (Scale).



80 member ensemble means


$N_e = 20$
$N_e = 40$
$N_e = 60$
$N_e = 80$
Free
ENKF
SEnT
Scale
PF
AEM
0.09
0.06
0.03
0.00
0.06
0.04
0.02
0.00
a_0 a_1 a_2 a_3 a_4 a_5
Ice Categories

Future directions:

- Extend to sea ice thickness observations
- Extend to the full, spatially-extended CICE model

Boden & Grooms, *Two-Step Ensemble Data Assimilation on the Simplex; Application to Sea Ice Concentration*, submitted 2025.

ian.grooms@colorado.edu

Simplex Ensemble Transport (SEnT):

- Step One: Update ensemble SIC (equiv. a_0)
- Step Two:
  - Use forecast ensemble to estimate parameters of joint mixed Dirichlet on $(a_0, \boldsymbol{a})$.
  - For each ensemble member (parallel):
    - Solve discrete, low-dim linear programming problem
    - Determine the member's analysis class.
    - If target and analysis class are the same, use scaling to update $\boldsymbol{a}$. Else
    - Transport forecast member to $[0,1]^{N_{\text{cat}}}$
    - Transport from $[0,1]^{N_{\text{cat}}}$ into the target analysis class

The ice in category i has mean thickness h_i which must lie between the bounds that define the category

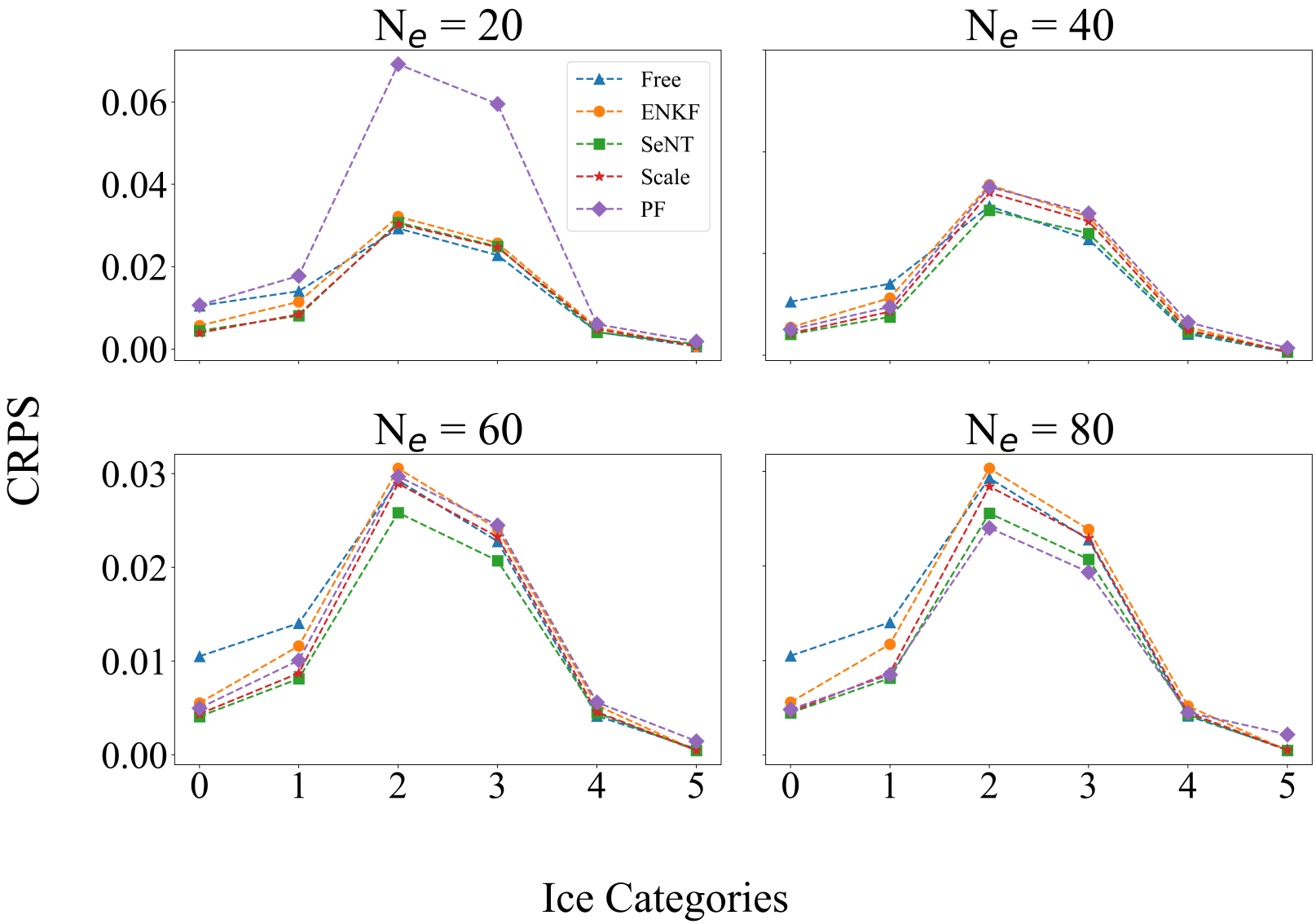
$$H_i < h_i < H_{i+1}.$$

We transform to an 'extended state' representation using

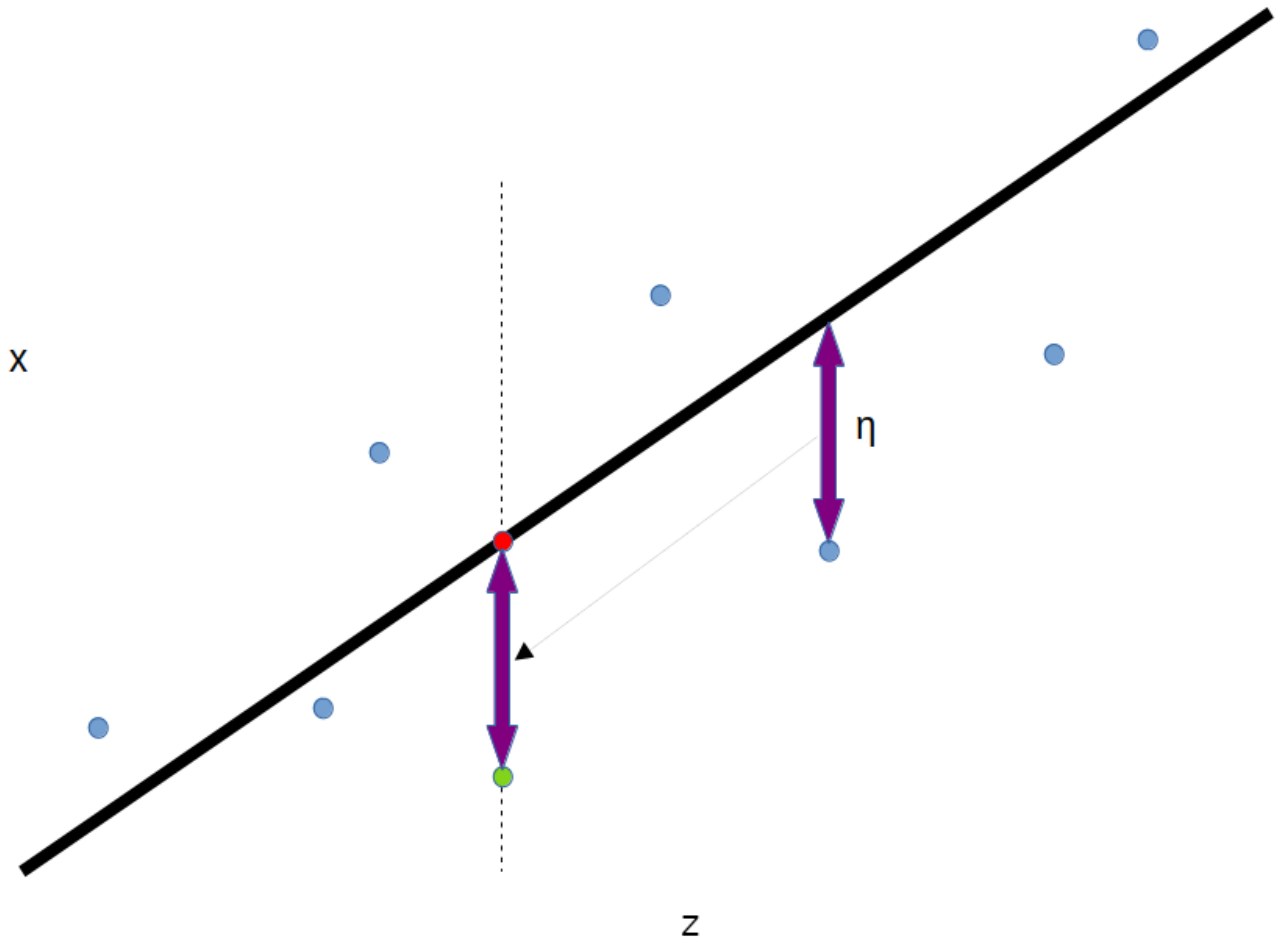
$$\begin{aligned} x_i^l + x_i^r &= a_i \\ H_i x_i^l + H_{i+1} x_i^r &= h_i a_i \end{aligned}$$

The extended state vector $a_0, x_1^l, x_1^r, \ldots, x_{N_{\text{cat}}}^l, x_{N_{\text{cat}}}^r$ is on a $2N_{\text{cat}} + 1$-dimensional simplex, and we apply SEnT there.

Scaling doesn't change thicknesses h_i; it only changes the areas a_i.


$N_e = 20$
$N_e = 40$
$N_e = 60$
$N_e = 80$
Free
ENKF
SeNT
Scale
PF
CRPS
0.06
0.04
0.02
0.00
0.03
0.02
0.01
0.00
0
1
2
3
4
5
Ice Categories

Conditional sampling of x_n^a given z_n^a via regression



$$x = \beta_0 + \beta_1 z + \eta$$