Novel Uncertainty Quantification Strategies for Passive Microwave-Sensed Sea Ice Concentration Retrievals for use in Non-gaussian Data Assimilation

Henry Santer

October 2025

A scholarly paper in partial fulfillment of the requirements for the degree of Master of Science

Department of Atmospheric and Oceanic Science, University of Maryland

College Park, Maryland

Research Advisor: Jonathan Poterjoy

1 Introduction

At a fundamental level, data assimilation (DA) is the process of combining prior information about the state of some dynamical system with observations to form a posterior estimate of the system's state. The primary goals are to a) construct the posterior state estimate in a way that respects the uncertainties of both the prior and the observations, and b) quantify the uncertainty surrounding the posterior estimate. When the latter condition is met, the posterior can be used as model *initial conditions* from which to initialize a forecast, establishing the basis for a *cycling data assimilation* system, where the posterior evolves forward in time via the system dynamics and becomes the prior for the next time observations are assimilated. Such cycling DA systems are routinely used at operational centers around the world for weather and climate prediction.

Data assimilation methods differ primarily in how they seek to represent the distributions of the prior, observations, and posterior. The Ensemble Kalman Filter (EnKF; Burgers et al., 1998) and its variants assume that each component follows a multivariate Gaussian distribution, where the observation error distribution has a constant mean μ and covariance matrix $\mathbf{R}$ and the background (prior) error covariance matrix is allowed to evolve in time according to the model dynamics. This assumption allows it to scale sufficiently to be computationally feasible for use with large geophysical models; however, it also makes these algorithms susceptible to errors when the distributions of the prior and observation errors are not actually Gaussian (Lei and Bickel, 2011; Santer et al., 2025). Particle filters (van Leeuwen et al., 2019), on the other hand, are more flexible DA methods that can accommodate arbitrary distributions, but which require prohibitively large ensemble sizes when observations are strongly informative over the entire domain of the state (see Section 2).

In recent years, a number of assimilation algorithms have been proposed which are simultaneously scalable and can accommodate non-Gaussian errors. This has been accomplished, for example, by applying localization to reduce the dimensionality limitations of particle filters (e.g., Poterjoy, 2022); by sequentially transforming particles from the prior to the posterior according to a particle flow induced by the posterior (Hu and van Leeuwen, 2021); and by more general representations of the prior distribution combined with Gaussian anamorphosis (Anderson, 2023). These methods are simultaneously computationally feasible with large state vectors and able to accommodate non-Gaussian errors. Likewise, these approaches allow for flexible choices of likelihood functions, which need to be specified in a way that is faithful to the real-world relationship between observations and the states being observed. This property of non-Gaussian ensemble filters is the primary topic of the present research.

Past efforts to estimate observation likelihoods during data assimilation have focused primarily on the case where observation errors are sufficiently characterized by their second moment. A common approach is to accumulate statistics of the background, analysis, and observations, and then use the innovation (observation minus background) and residual (analysis minus background) statistics to estimate **R**. The Desroziers et al. (2005) "consistency check" method calculates **R** as the expectation of the product of innovations and residuals, which in practice can be found by computing sample averages over previous data assimilation cycles. Similarly, the methods of Karspeck (2016) and Santer et al. (2025) combine innovation statistics with ensemble-estimated background covariances to estimate scalar observation error variances. Each of these methods leverages the fact that the innovation variance can be decomposed into the observation error variance and the background error variance, and each has been shown to be effective when the true unknown observation likelihoods are Gaussian and unbiased. For the case where likelihoods are expected to be non-Gaussian, Hu et al. (2024) presents a method for nonparametric estimation of the observation error pdf, based on taking a deconvolution of the background error pdf from the total innovation pdf. This method is straightforward to incorporate into particle filter methods and has also been implemented into variational frameworks (**?**). However, it presents its estimates in terms of an unknown observation error, and so still requires computation of *forward operators* that transform model states into observation space, which can be expensive and may introduce additional sources of error. The method proposed in this paper estimates likelihood functions directly, so that data assimilation strategies that allow for flexible choices of likelihood function (e.g., particle filters) can completely forego the need for a forward operator.

Sea ice is an important component of the Earth system because of its impact on the global energy balance, leading it to influence the atmosphere and ocean on both short and long timescales. Many of the variables that are modeled and observed to constrain sea ice are bounded. Among these variables is *sea ice concentration* (SIC), which is the fractional coverage of sea ice within a given observing area. A SIC of 0% corresponds to zero ice coverage, and a SIC of 100% corresponds to full ice coverage. As a result of this boundedness, observation errors for observations of SIC are highly state-dependent and also non-Gaussian. For example, when the true sea ice concentration is very close to 100%, observation errors must either be negative or very slightly positive, because more positive observation errors would lead to values exceeding 100%. The likelihood function for SIC observations, therefore, is both non-Gaussian and state dependent, making SIC observations an ideal candidate for investigating the benefits of the above non-Gaussian data assimilation techniques. The objective of this paper is to create fully nonparametric estimates of the non-Gaussian distributions of SIC observations given model states to facilitate their assimilation in fully non-Gaussian settings. In section 4, we introduce a novel technique for estimating state-dependent, non-Gaussian likelihood functions based on training data that can be accumulated during DA.

The rest of this manuscript is organized as follows. Section 2 provides an overview of Bayesian data assimilation and demonstrates the value of specifying non-Gaussian likelihoods when using DA methods that can accommodate them. Section 3 describes the passive microwave-sensed SIC observations and the NASA Team 2 retrieval algorithm that generates them. Section 4 introduces a non-parametric likelihood estimation method based on kernel embeddings of conditional distributions. Section 5 presents results applying the method to assimilation experiments with idealized models where the true likelihoods are known, and Section 6 then shows likelihood estimates for real SIC observations. Conclusions and future research directions are discussed in Section 7.

2 Bayesian Filtering

Consider a random vector $\mathbf{x}$ representing all of the prognostic state variables of a discretized model of some system. Let $\mathbf{y}$ be a vector containing observations of that system; these observations are associated with the true state $\mathbf{x}^{\mathbf{t}}$ via a measurement operator h and an observation error ϵ as follows:

$$\mathbf{y} = h[\mathbf{x}^{\mathbf{t}}] + \epsilon. \tag{1}$$

The observation error ϵ is treated as a random variable, and so $\mathbf{y}$ is also a random variable. In practice, we obtain an actual measurement $\mathbf{y}^{\mathbf{o}}$ of the system, and then seek an estimate of the posterior probability density $p(\mathbf{x}|\mathbf{y}^{\mathbf{o}})$ of $\mathbf{x}$ conditioned on $\mathbf{y}^{\mathbf{o}}$. Given a prior probability $p(\mathbf{x})$ describing initial beliefs about $\mathbf{x}$ and the likelihood $p(\mathbf{y}^{\mathbf{o}}|\mathbf{x})$ of model states given the observation, *Bayes' Theorem* shows that the posterior probability can be written as

$$p(\mathbf{x}|\mathbf{y}^{\mathbf{o}}) \propto p(\mathbf{y}^{\mathbf{o}}|\mathbf{x})p(\mathbf{x}). \tag{2}$$

Multiple distinct data assimilation algorithms follow from this result, depending on how the prior and likelihood are specified. One very general method is *particle filtering.* The general particle filtering strategy is to model the posterior through a Monte Carlo approach. Consider a set $\{\mathbf{x_n}\}_{n=1}^{N_e}$ of states sampled from $p(\mathbf{x})$. An approximation of $p(\mathbf{x})$ can be constructed with a sum of delta functions centered at each particle:

$$p(\mathbf{x}) \approx \frac{1}{N_e}\sum_{n=1}^{N_e} \delta(\mathbf{x} - \mathbf{x_n}). \tag{3}$$

The posterior density can then be approximated with a weighted sum of the same delta functions, where the weights w_n are determined based on the likelihood. In particular,

$$w_n = p(\mathbf{y^o}|\mathbf{x_n}), \tag{4}$$

$$W = \sum_{n=1}^{N_e} w_n, \tag{5}$$

$$p(\mathbf{x}|\mathbf{y^o}) \approx \sum_{n=1}^{N_e} \frac{w_n}{W}\delta(\mathbf{x} - \mathbf{x_n}), \tag{6}$$

from which other useful statistics of the posterior can also be approximated (e.g., the posterior mean and variance). Equations (3)-(6) form the basis of the basic sequential importance sampling particle filter when the prior is used as the proposal density and the weights from the previous filtering step are assumed to be equal. This filter makes no assumptions about the forms of the prior, likelihood, and posterior in order to function. However, the ensemble size N_e required to prevent the collapse of particle weights is generally computationally infeasible for moderate-to-large dimensional applications. Snyder et al. (2008) show that, for the case when the prior and observation errors are Gaussian, N_e must increase exponentially with the variance of the observation log-likelihood in order to prevent weight collapse. In practice, the variance of the observation log-likelihood acts as a measure of the "degrees of freedom" of the state with respect to the observations; for example, weight collapse is likely to occur when attempting to assimilate a dense network of accurate observations with independent errors, but may not occur when only a small number of accurate observations is assimilated, or when many observations with large or strongly correlated errors are assimilated.

To overcome these challenges, the present study uses the regularized local particle filter described in Poterjoy (2022), which is sufficiently scalable for geophysical models while retaining the capacity for arbitrary priors and likelihoods. The primary innovation is to expand each scalar weight w_n into a vector $\omega_\mathbf{n}$ with the same length as the state vector. The vector weights are computed from the scalar weights and a localization function that depends on the Euclidean distance between the observation and each model grid point. Observations are then assimilated sequentially as follows. First, prior particles are resampled according to the scalar weights; let $\{\mathbf{x_{k_n}}\}_{n=1}^{N_e}$ be the resampled particles. Then, the original prior particles are merged with the resampled particles to yield particles conditioned on each observation up to the current scalar observation y:

$$\mathbf{x_{n,y}} = \mathbf{\bar{x}_y} + \mathbf{r_1} \circ (\mathbf{x_{k_n}} - \mathbf{\bar{x}_y}) + \mathbf{r_2} \circ (\mathbf{x^n} - \mathbf{\bar{x}_y}), \tag{7}$$

where $\mathbf{\bar{x}_y}$ is the posterior mean calculated from the localized (vector) weights for all observations up to y and $\circ$ represents the Schur (i.e., elementwise) product. The vector coefficients $\mathbf{r_1}$ and $\mathbf{r_2}$ are chosen so that $\mathbf{x_{n,y}}$ resembles samples from the standard particle filter posterior for variables at grid points that are close to the location of observations, but resembles the prior particles further away, where observations are assumed to have less influence. Finally, a regularization is applied to particle weights to stabilize the filter and prevent particle collapse when assimilating dense networks of accurate observations. See Poterjoy (2022) and Poterjoy et al. (2019) for a more thorough derivation of the update equations and an outline of the filter algorithm.

The regularized local particle filter provides a usable tool for assimilating observations with non-Gaussian likelihoods. We next provide an overview of the SIC observations and of some initial efforts to quantify their uncertainty, before describing techniques for estimating their full distributions with statistics generated during data assimilation.

3 Observations of Sea Ice Concentration

Satellite remote sensing of SIC is predicated on differences in brightness temperature of sea ice and open water at different polarizations and microwave frequencies. Satellite-based microwave radiometers receive radiances that are a composite of the surface radiation, atmospheric radiation, and background radiation in space. Cavalieri et al. (1984) show that, in the polar regions, the atmospheric and space components of the total composite radiation can usually be neglected, so that the measured brightness temperature T_B can be considered equal to the surface microwave emission alone. In the polar Arctic, the surface microwave emission can then be decomposed into emission from open water and emission from ice:

$$T_B \approx T_w(1 - C) + T_{ICE}C, \tag{8}$$

where T_w and T_{ICE} are the radiances of the open ocean and ice, respectively, and C is the sea ice concentration. Microwave radiances can also be used to distinguish first-year ice from multi-year ice, which have different radiative properties due to the abundance of brine present in first-year ice (Comiso, 2005). The ice radiance can therefore be further decomposed so that

$$T_B \approx T_w(1 - C) + T_{FYI}C(1 - F) + T_{MYI}CF, \tag{9}$$

where T_{FYI} and T_{MYI} are the radiances of the first year and multi-year ice and F is the multi-year ice fraction. The goal of most SIC retrieval algorithms is to establish characteristic values of T_w, T_{FYI}, and T_{MYI} at different frequencies for both vertical and horizontal polarizations, and then to use equation (9) to find the values of C and F that reconstruct the observed radiances as closely as possible. The chosen values of T_w, T_{FYI}, and T_{MYI} act as parameters for these algorithms and are called *tie points*; tie points determine how the measured radiances are used to retrieve SIC and therefore directly influence the manner in which observation errors depend on the true ice concentration.

3.1 The NASA Team 2 Retrieval Algorithm

Although atmospheric radiances are often neglected for the purposes of designing SIC algorithms, weather can still introduce uncertainties into SIC retrievals. Over the open ocean, surface winds act to increase brightness temperatures by roughening the ocean surface and mixing the cooler skin layer with warmer water below it, especially for horizontal polarizations (Zappa et al., 2019; Meissner and Wentz, 2012). Cloud liquid water can also contribute to increased microwave emission, and can increase computed ice concentrations by as much as 8% (Cavalieri et al., 1984).

The NASA Team 2 Algorithm (Markus and Cavalieri, 2009) uses three different microwave channels (19GHz, 37GHz, 89GHz) to compute SIC and additionally incorporates a forward atmospheric radiative transfer model and an additional weather filter to account for the above atmospheric effects. The algorithm proceeds as follows. First, brightness temperatures are computed for each polarization p and each of the three frequencies ν given different sea ice concentrations. This computation is done for each of 12 different atmospheric profiles $\{P_w\}_{w=1}^{12}$, and for two different types of ice C_A and C_C, where C_A is the concentration of both first year and multi-year ice and C_C is the concentration of ice with significant surface glazing, which has different emissive properties. C_A and C_C are considered in increments of 1%, so the result of these computations is a lookup table of dimension (101 x 101 x 12) for each p and ν. Next, define the *spectral gradient ratio* GR and *polarization ratio* PR as follows:

$$PR(\nu) = \frac{TB(\nu V) - TB(\nu H)}{TB(\nu V) + TB(\nu H)}, \tag{10}$$

$$GR(\nu_1 p, \nu_2 p) = \frac{TB(\nu_1 p) - TB(\nu_2 p)}{TB(\nu_1 p) + TB(\nu_2 p)}. \tag{11}$$

From the computed brightness temperatures, three more (101 x 101 x 12) tables are constructed:

$$LUT_{PR19}(C_A, C_C, w) = GR(37V, 19V)\sin(\phi_{19}) + PR(19)\cos(\phi_{19}) \tag{12}$$

$$LUT_{PR89}(C_A, C_C, w) = GR(37V, 19V)\sin(\phi_{89}) + PR(89)\cos(\phi_{89}) \tag{13}$$

$$LUT_{\Delta GR}(C_A, C_C, w) = GR(89H, 19H) + GR(89V, 19V), \tag{14}$$

where the angles ϕ_{19} and ϕ_{89} are chosen to simplify calculations in the algorithm -

see Markus and Cavalieri (2009) for details. The same three values are computed with the actual observed brightness temperatures, and the triple (C_A^*, C_C^*, w^*) with the best least-squares fit to the observed values is chosen as the SIC retrieval. The final SIC is calculated as $C = C_A^* + C_C^*$.

Two additional filters are implemented to further mitigate contamination of SIC retrievals that result from weather effects. The first is based on monthly climatological sea surface temperatures (SSTs). Any pixels in the northern hemisphere with a monthly average SST greater than 278K ($= 4.85°C$) are set to have SIC=0 throughout the month. The second filter is based on radiative transfer model experiments with different prescribed atmospheric and oceanic conditions. Gloersen and Cavalieri (1986) found that atmospheric contributions to open-water microwave retrievals were most significant when the gradient ratio $GR(37V, 18V)$ exceeded 0.08. Accordingly, they suggest placing an upper bound on $GR(37V, 18V)$ to suppress spurious ice over open ocean due to weather effects, beyond which the retrieved ice concentration is set to 0. The exact choice of upper bound varies with the observing instrument; AMSR-2 retrievals currently set an upper bound of 0.05 on $GR(37V, 19V)$, and an upper bound of 0.045 on $GR(22V, 19V)$. The GR weather filters eliminate the majority of severe weather effects over the open ocean. However, they also have the potential to erroneously set small (≤15%) SIC retrievals to exactly 0%, even when some ice is actually present.

The NASA Team 2 algorithm's performance has been verified against multiple sources and observing systems, and it has been found to generally produce retrievals that are within operational uncertainty requirements (Meier et al., 2017). However, no quantification of its uncertainty has been done that captures the inherent non-Gaussianity of SIC retrievals, and errors tend to be larger closer to the ice edge when SIC is closer to 0. We also note that the weather filter and SST filter still introduce additional nonlinearity into the SIC retrievals despite their benefits, further complicating attempts to model their uncertainty. As of this writing, there is an outstanding need for uncertainty quantification techniques for SIC retrievals that respect both their state dependence and their non-Gaussianity.

4 Nonparametric Estimation of $p(\mathbf{y}|\mathbf{x})$ with Kernel Embeddings of Conditional Distributions

Berry and Harlim (2017) (hereafter BH17) estimate observation forward operator errors during data assimilation with a secondary filter that can be coupled with multiple data assimilation schemes. In doing so, they form nonparametric estimates of probability densities $p(\mathbf{y}|\mathbf{b})$ for observations $\mathbf{y}$ given an observed forward operator error $\mathbf{b}$. We now adapt their methodology to directly estimate the conditional probability densities $p(\mathbf{y}|\mathbf{x})$ for different values of $\mathbf{x}$, which in turn can be used to calculate the likelihood $p(\mathbf{y^o}|\mathbf{x})$ of of model states given a specific observation.

Let $\{\mathbf{x_i}\}_{i=1}^{N} \in \mathcal{X} \subset \mathbb{R}^n$ be a dataset of model state estimates with sampling density $q(\mathbf{x})$ over the manifold $\mathcal{X}$. The first step is to associate $\{\mathbf{x_i}\}$ with a Hilbert space of functions from $\mathcal{X}$ to $\mathbb{R}$. Hilbert spaces are vector spaces with additional desirable properties, so that this allows us to use techniques from linear algebra on functions of $\{\mathbf{x_i}\}$. As in BH17, we accomplish this using the Diffusion Maps algorithm (Coifman and Lafon, 2006). The diffusion maps algorithm applied to $\{\mathbf{x_i}\}$ constructs an approximation to the weighed Laplacian operator $\mathscr{L} = \nabla \log(q) \cdot \nabla + \Delta$, which describes diffusion along $\mathcal{X}$ weighted by the density p. Notably, Coifman and Lafon (2006) show that the eigenfunctions of $\mathscr{L}$ form an orthonormal basis $\{\varphi_k\}$ for the Hilbert space $L^2(\mathbb{R}, q)$ (i.e. of square-integrable functions with respect to the density q). Practically speaking, the diffusion maps representation of $\{\mathbf{x_i}\}$ can be computed by considering a random walk on the dataset. Transition probabilities are given by a Gaussian kernel (distinct from the kernel functions discussed below) so that they are large for close-together points and small for distant points. We note that the bandwidth of this kernel is a tunable parameter of this method, and explore its impact on the estimated likelihood functions in Section 5. We also employ a k-nearest neighbors (kNN) heuristic to eliminate transition probabilities for far-away points and reduce computational costs.

The eigenvectors $\{\boldsymbol{\varphi}_k\}$ of the corresponding Markov matrix represent the basis functions evaluated at each data point (e.g., the first entry of $\boldsymbol{\varphi}_3$ corresponds to $\varphi_3(\mathbf{x_1})$). Based on this eigenfunction representation, define the *kernel function*

$$K(\mathbf{x_1}, \mathbf{x_2}) = \sum_i \varphi_i(\mathbf{x_1})\varphi_i(\mathbf{x_2}). \tag{15}$$

For any $\mathbf{x} \in \mathcal{X}$, we can define the function $K_{\mathbf{x}} \coloneqq K(\mathbf{x}, \cdot)$. For any function $f \in L^2(\mathbb{R}, q)$, then, we have

$$\begin{aligned} \langle\, K_{\mathbf{x}}, f \,\rangle_q &= \left\langle\, \sum_i \varphi_i(\mathbf{x})\varphi_i(\cdot)\;, \sum_i f_i \varphi_i(\cdot) \right\rangle_q , && (16) \\ &= \sum_i \varphi_i(\mathbf{x}) \left\langle\; \varphi_i(\cdot)\;, \sum_i f_i \varphi_i(\cdot) \right\rangle_q , && (17) \\ &= \sum_i \varphi_i(\mathbf{x}) f_i && (18) \\ &= f(\mathbf{x}) && (19) \end{aligned}$$

where we use the orthonormality of the eigenfunctions for the third equality, and the subscript q represents the use of the q-weighted inner product instead of the standard inner product on $L^2(\mathbb{R})$. This is the so-called *reproducing property*, and shows that $L^2(\mathbb{R}, q)$ is a *reproducing kernel Hilbert space* with reproducing kernel K (RKHS; Aronszajn, 1950). RKHS's have wide applications because they guarantee that evaluations of any function can always be computed as an inner product with the reproducing kernel, so that involved manipulations of functions reduce to operations on matrices of kernel evaluations. We repeat the above computations and reasoning to compute a diffusion map representation for the dataset $\{\mathbf{y_i}\}_{i=1}^N \in \mathcal{Y} \subset \mathbb{R}^m$ of observations corresponding to the model states, with sampling density $\tilde{q}$. This results in an orthonormal basis of functions $\{\phi_k\}$ for the RKHS $L^2(\mathbb{R}, \tilde{q})$ with reproducing kernel $\tilde{K}$.

The RKHS machinery can be used to estimate conditional densities using the theory of kernel embeddings of conditional distributions (KECD; Song et al., 2013). Consider a random variable $\mathbf{X}$ that takes values in $\mathcal{X}$ with distribution $p(\mathbf{X})$. The *kernel mean embedding* of $p(\mathbf{X})$ into $L^2(\mathbb{R}, q)$ is defined as

$$\mu_{\mathbf{X}} = \mathbb{E}[K(\mathbf{X}, \cdot)] = \int_{\mathcal{X}} K(\mathbf{X}, \cdot) dp(\mathbf{x}). \tag{20}$$

It follows from the reproducing property that $\mu_{\mathbf{X}}$ encodes expectations with respect to $p(\mathbf{X})$ as inner products in the RKHS; specifically, we have $\mathbb{E}_{\mathbf{X}}[f(\mathbf{X})] = \langle\, f, \mu_{\mathbf{X}} \,\rangle_q$. We can also encode joint distributions in a similar way. Let $\mathbf{Y}$ be a random variable over $\mathcal{Y}$. The joint distribution $p(\mathbf{X}, \mathbf{Y})$ can be embedded into the tensor product space $L^2(\mathbb{R}, q) \otimes L^2(\mathbb{R}, q)$ as

$$\mathscr{C}_{\mathbf{XY}} \coloneqq \int_{\mathcal{X}\times\mathcal{Y}} K(\mathbf{x}, \cdot) \otimes \tilde{K}(\mathbf{y}, \cdot) dp(\mathbf{x}, \mathbf{y}). \tag{21}$$

Similarly to $\mu_{\mathbf{X}}$, we can use $\mathscr{C}_{\mathbf{XY}}$ to evaluate the expectation

$$\mathbb{E}_{\mathbf{XY}}[f(\mathbf{X})g(\mathbf{Y})] = \langle f \otimes g \,, \mathscr{C}_{\mathbf{XY}} \rangle_{q\times\tilde{q}} = \langle f \,, \mathscr{C}_{\mathbf{XY}} g\tilde{q} \rangle_q \,, \tag{22}$$

where in the second equality we used the standard equivalence of tensors with linear maps (e.g., Winitzki (2010)). Equation (22) says that $\mathscr{C}_{\mathbf{XY}}$ can be used to compute the cross-covariance of two functions $f \in L^2(\mathbb{R}, q)$ and $g \in L^2(\mathbb{R}, \tilde{q})$ as an inner product. Finally, we define the kernel embedding of the conditional distribution $p(\mathbf{Y}|\mathbf{X})$:

$$\mu_{\mathbf{Y}|\mathbf{x}} = \int_{\mathcal{Y}} \tilde{K}(\mathbf{y}, \cdot) dp(\mathbf{y}|\mathbf{x}) \tag{23}$$

which satisfies $\mathbb{E}_{\mathbf{Y}|\mathbf{x}}[g(\mathbf{Y})] = \left\langle g, \mu_{\mathbf{Y}|\mathbf{x}} \right\rangle_{\tilde{q}}$. The conditional embedding represents a family of functions in the observation RKHS, indexed by the conditioning variable $\mathbf{x}$. BH17 show that the relationship between $\mathbf{x}$ and $\mu_{\mathbf{Y}|\mathbf{x}}$ is given by

$$\mu_{\mathbf{Y}|\mathbf{x}} \approx q\mathscr{C}_{\mathbf{YX}}\mathscr{C}_{\mathbf{XX}}^{-1} K(\mathbf{x}, \cdot) \tag{24}$$

where the approximation can be considered exact under certain regularity assumptions.

So far, we have presented the diffusion maps algorithm and the kernel function using an arbitrarily large (potentially infinite) number of eigenfunctions φ_k and ϕ_k. BH17 shows that the conditional embedding can be used to evaluate conditional densities directly if only a finite collection of eigenfunctions is retained from the diffusion maps representation and used to construct approximations of the kernel mean embeddings. Keeping only the M eigenfunctions with the largest eigenvalues, we can write the conditional probability of observation $\mathbf{y_i}$ given model state $\mathbf{x_j}$ as

$$p(\mathbf{y_i}|\mathbf{x_j}) = \sum_{k=1}^{M} \mu_{\mathbf{Y}|\mathbf{x}_j,k} \phi_k(\mathbf{y_i}) \tilde{q}(\mathbf{y_i}), \tag{25}$$

where the coefficients $\mu_{\mathbf{Y}|\mathbf{x}_j,k} = \left\langle \mu_{\mathbf{Y}|\mathbf{x_j}}, \phi_k \right\rangle_{\tilde{q}}$ can be approximated as

$$\mu_{\mathbf{Y}|\mathbf{x_j},\mathbf{k}} = \sum_{l=1}^{M} \varphi_l(\mathbf{x_j}) [\mathbf{C_{YX}} \mathbf{C_{XX}^{-1}}]_{kl}, \tag{26}$$

$$[\mathbf{C_{YX}}]_{lk} = \frac{1}{N} \sum_{n=1}^{N} \phi_l(\mathbf{y_n}) \varphi_k(\mathbf{x_n}), \tag{27}$$

$$[\mathbf{C_{XX}}]_{lk} = \frac{1}{N} \sum_{n=1}^{N} \varphi_l(\mathbf{y_n}) \varphi_k(\mathbf{x_n}). \tag{28}$$

Equations (26) represent a sample estimate of the conditional embedding when only a finite number of eigenfunctions is used. In practice, we apply these equations by collecting matched sets of model states and observations to use as training data and using them to create an $N \times N$ matrix $\mathbf{L}$ whose (i, j)th entry is $p(\mathbf{y_i}|\mathbf{x_j})$. Given a new observation-state pair $(\mathbf{x}^*, \mathbf{y}^*)$ from outside the training data (e.g., from a later data assimilation cycle), we then use the Nyström extension (Nyström, 1930) to map those points into their respective RKHS's. We then identify the points from the training data closest to the extrapolated points (where distance is measured by the RKHS norms), and choose the corresponding value from $\mathbf{L}$ as the likelihood $p(\mathbf{y}^*|\mathbf{x}^*)$. This likelihood can then be used directly to weight particles with the LPF.

The above method is nonparametric in the sense that it assumes no form of the underlying distribution of the observations or model states. The computational cost is generally dominated by the eigendecomposition of the Markov matrices generated during the diffusion maps representation, which scales primarily with the size of the training data. However, we can apply a k-nearest-neighbors (kNN) heuristic during the construction of the Markov matrices, where only a set number of points closest to each data point are assigned nonzero transition probabilities. This sparsens the matrices, which enables faster computations, and can also reduce spurious transition probabilities across distant data points, in a manner similar to how covariance localization reduces spurious covariances in ensemble data assimilation.

5 Idealized Model Experiments

We begin evaluating the KECD approach by performing a collection of experiments with the Lorenz (1995) 40-variable model, which represents the behavior of some atmospheric quantity along a latitude circle and has terms representing advection, internal dissipation, and external forcing. The governing equations for the Lorenz 96 model are

$$\frac{dX_k}{dt} = (X_{k+1} - X_{k-2})X_{k-1} = X_k + F \tag{29}$$

for $k \in \{1, 2, ..., 40\}$. X_{-1}, X_0, and X_{41} are taken to be equal to X_{39}, X_{40}, and X_1, respectively, so that the equations are defined for each k and represent behavior on a closed circle.

We spin up an 80-member ensemble for 100 time steps ($dt = 0.05$) and then perform data assimilation every 10 time steps for 2500 cycles. A single model run is generated to simulate the true state; initial conditions for both the ensemble and the truth run are generated by creating state vectors from random noise, so they can be considered to arise from the same climatological distribution. Observations of the system are generated by adding perturbations from known distributions to the truth run. They are then at every other grid point (i.e., 20 observations per cycle). All experiments use the Poterjoy (2022) regularized local particle filter to assimilate observations. For each observation error distribution, we perform a data assimilation experiment a) with likelihood functions estimated from KECD; b) assuming observation errors are Gaussian with mean zero and variance 1; and c) with likelihood functions specified assuming perfect knowledge of the observation error distribution. For the training data, we use observations from the 400 most recent assimilation cycles (i.e., the most recent 8000 observations). The corresponding model states are random draws from the posterior distribution valid at the same time, which are used as a proxy for the true state. After the first time the KECD estimation is performed, we repeat it every 50 assimilation cycles, using the most recently estimated likelihood functions in the meantime. Unless otherwise stated, a bandwidth of 0.02 is chosen for the Gaussian kernel used in the diffusion maps algorithm, and for each data point, we only retain nonzero transition probabilities for the closest 800 points. We represent the RKHSs using the 50 eigenfunctions with the largest eigenvalues for each space.

Figure 1 shows the final estimated conditional distributions $p(\mathbf{y}|\mathbf{x})$ for different values of $\mathbf{x}$ (left column) and analysis-minus-truth root mean square errors (right column) for two different types of observation error. We show conditional distributions for the 10th, 25th, 50th, 75th, and 90th percentiles of model states. The first row corresponds to a basic case where observation errors are Gaussian with variance 1, but with a state-dependent bias: observations of model states greater than 0 have a bias of 2, while observations less than 0 have a bias of -2. After training and an adjustment period of about 500 cycles, the KECD estimated distributions approximate the correct distributions well enough to match the performance of the ideal case where the true likelihood functions are known. We observe specifically that the state dependence of the conditional distributions is well captured: when $\mathbf{x} \approx 0$, the estimated distribution for $\mathbf{y}$ is bimodal, with modes that represent the two possible biases in observations. The second row shows experiments using observation errors following a Cauchy distribution with a scale parameter equal to 0.2, which is centered at 0 but has heavier tails than a Gaussian. In this case, KECD is still able to estimate conditional distributions that improve forecasts, but not to the extent that they match the skill of assimilating with the true likelihood. We also note that the estimated distributions vary considerably in shape with the state, despite the lack of state dependence in the true observation errors. In particular, KECD estimates more correct distributions for model states that are more central to the marginal distribution for $\mathbf{x}$.

5.0.1 Sensitivity to Parameter Tuning

The performance of the KECD method depends on the tuning of the bandwidth and other heuristics applied during the diffusion maps algorithm. Figure 2 shows the experiment with Cauchy random errors for three different configurations of the bandwidth and k-nearest-neighbors parameters. The middle row of Figure 2 is the same as the bottom row of Figure 1. The first row reduces both the bandwidth and the number of nearest neighbors by a factor of 2. The result is that the estimated conditional distributions are more homogenized with respect to the conditioning state, which is more reflective of the true likelihood functions and accordingly leads to improved forecast outcomes. In contrast, doubling the bandwidth and nearest neighbors (third row) leads to estimated conditional distributions that are considerably more spread out than the true distributions. The quality of the estimates toward the extremes of the marginal distribution for the state is also further degraded; most of the distributions no longer have modes at their corresponding model states. Using these distributions, the ensemble analyses are degraded to the point of being worse than when Gaussian likelihoods are used. In the context of the diffusion maps algorithm, larger values of both parameters correspond to faster diffusion to far away points. Although this will have a disproportionately large effect for heavy tailed distributions, we note that it is important in general to carefully specify these parameters for KECD, and that in this work we have determined reasonable values experimentally.

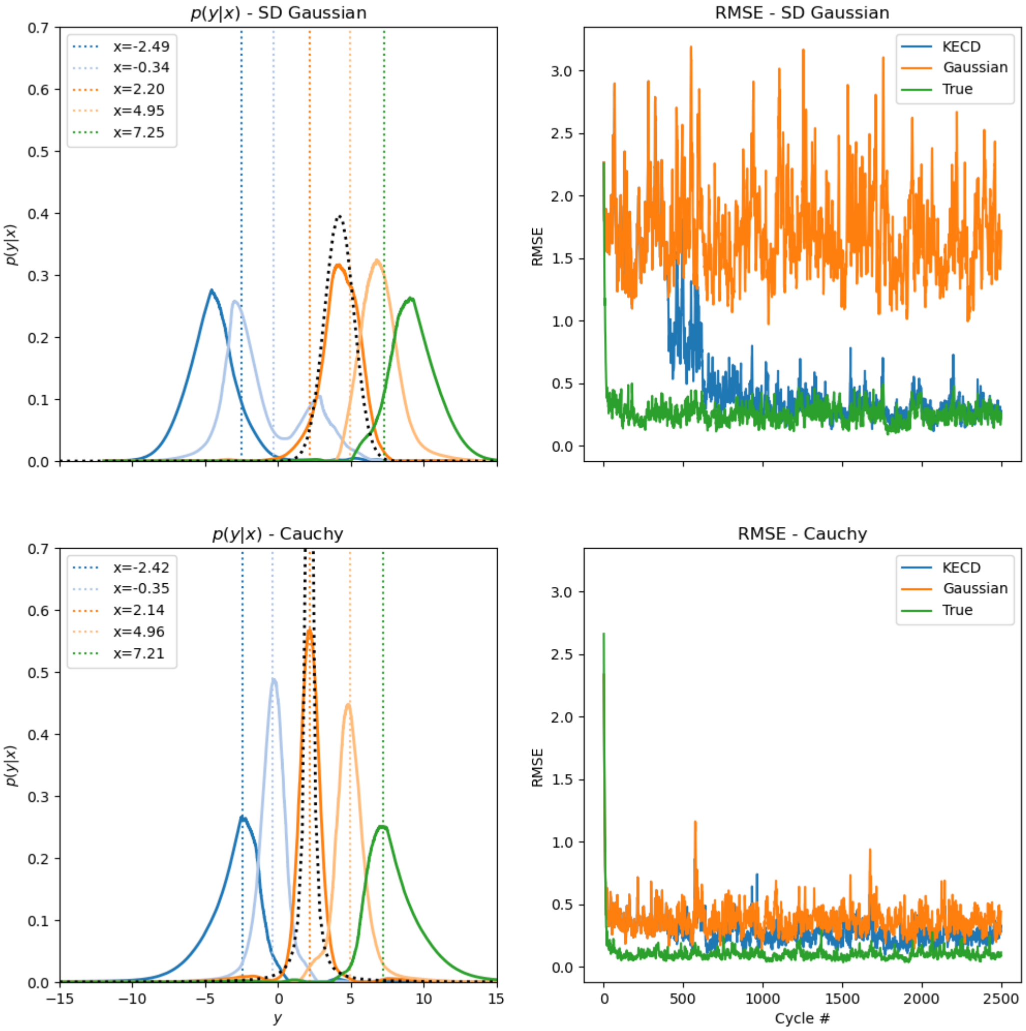

Figure 1: Estimates of $p(\mathbf{y}|\mathbf{x})$ for various SIC analysis values (left) and root mean square errors of analysis with respect to the true state (right) for experiments where observation errors are (top) Gaussian with a state dependent bias and (bottom) Cauchy distributed with location parameter 0 and scale parameter (0.2). Dashed vertical lines indicate the analysis states on which the estimated distributions are conditioned. The dotted black curve shows the true conditional distribution for the 50th percentile state.

6 Application to NASA Team 2 SIC Retrievals

Having demonstrated the potential benefits of the data-driven likelihood estimation method, we now apply it to training data collected from the Joint Center for Satellite Data Assimilation (JEDI) Sea-Ice Ocean and Coupled Assimilation (SOCA) DA system. The model states are SIC analyses from coupled atmosphere-ocean-sea ice

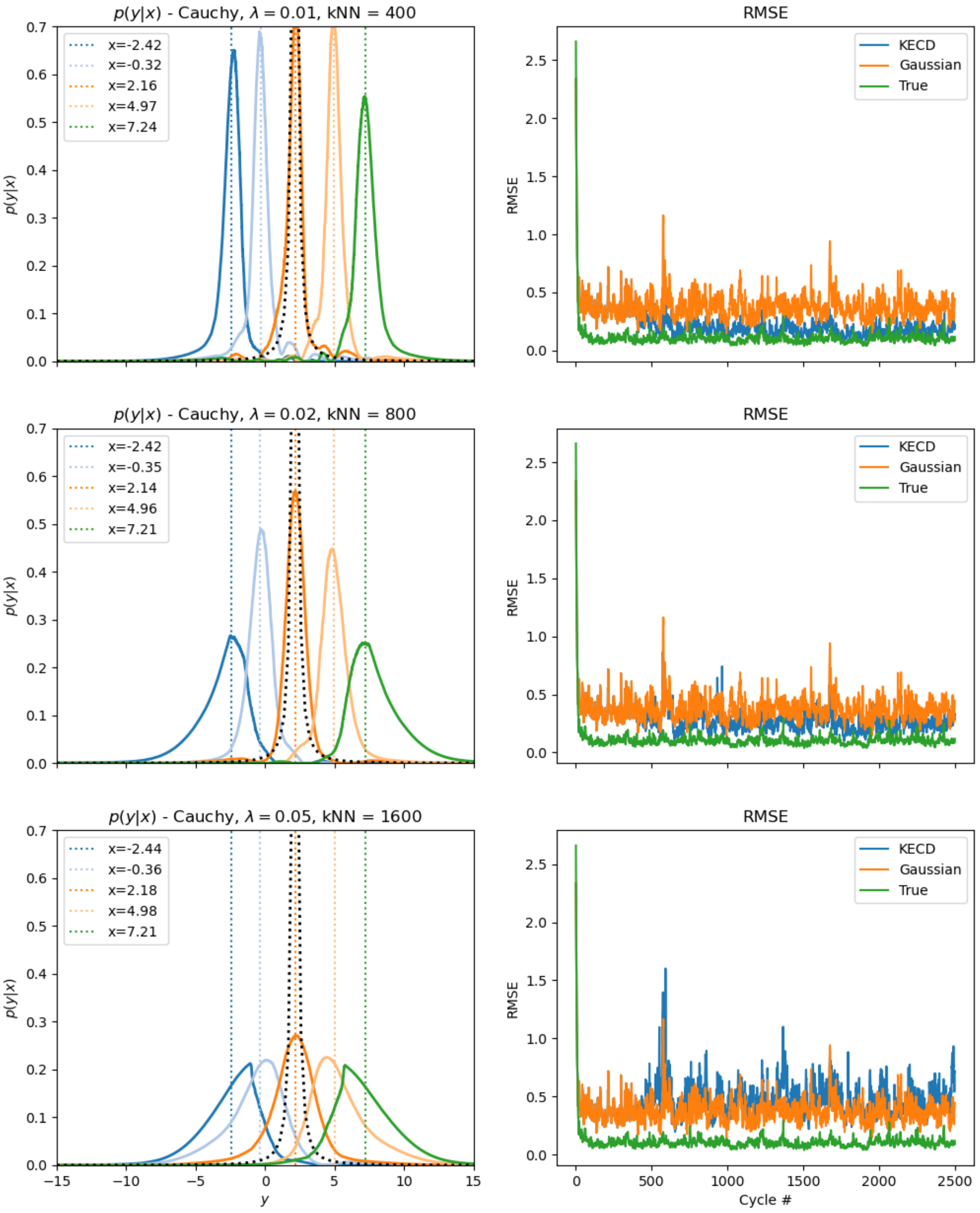

Figure 2: As in Figure (1), but with estimates of $p(\mathbf{y}|\mathbf{x})$ for Cauchy distributed observation errors as the bandwidth and k-nearest neighbors parameters vary from small (top) to large (bottom).

experiments, where the sea ice component is the Los Alamos Sea Ice Model (CICE) version 6 and the data assimilation system is the JEDI 3DVar scheme. Observations are ice concentrations calculated with the NASA Team 2 algorithm from brightness temperatures gathered by the Advanced Microwave Scanning Radiometer 2 (AMSR2), and are assimilated assuming Gaussian observation errors with variance 0.1. Figure 3 shows the distribution of SIC analysis states and observations valid at 00 UTC on August 1, 2021. The distribution of both the analysis states and observations are bimodal, and the total number of available valid training pairs is approximately 1.6×10^5; to make the eigendecompositions required for the diffusion maps algorithm computationally feasible, we choose a subset of 2×10^4 pairs to use for training. The pairs are sampled from the training data so that the distribution of analysis states in the training data is approximately uniform. This is done to ensure that the learned representations of the training datasets are not dominated by the sampling density; Coifman and Lafon (2006) present additional strategies for controlling the influence of the sampling density on the diffusion maps representation.

Figure 4 shows the SIC analysis field used for training (left) and observation PDF estimates for three different values of the SIC state (right). KECD estimates of $p(\mathbf{y}|\mathbf{x})$ respect both the state dependence and the boundedness of SIC observation errors. When SIC $= 0.86$, the observation PDF is highest between 0.9 and 1.0 and is negatively skewed. There is a considerable deviation from Gaussianity that would make assimilation with Gaussian likelihoods unsuitable. When SIC ≈ 0.5, the PDF is approximately Gaussian, with a spread consistent with the operationally prescribed variance. Close to the ice edge where SIC is small, $p(\mathbf{y}|\mathbf{x})$ is also highly non-Gaussian, with a large mode at 0 and multiple smaller modes. Despite the mode at 0, there is zero likelihood for observations between 0 and approximately 0.1. We attribute this to the weather filters employed by the Team 2 algorithm.

7 Discussion

In the idealized experiments, the likelihood functions estimated with the KECD method respect the non-Gaussianity and state dependence of the prescribed observa-

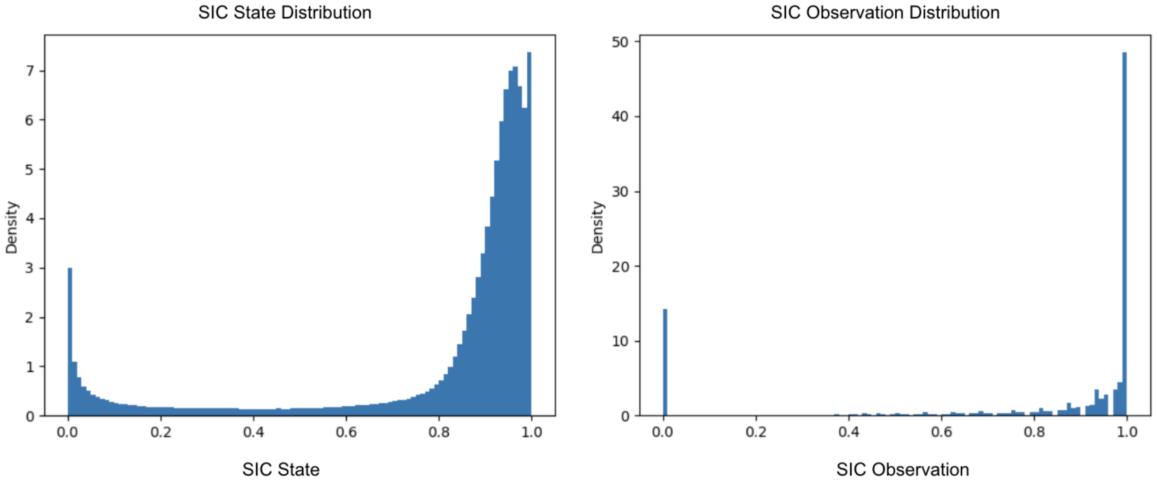

Figure 3: Distributions of model states and observations for SIC valid at 00 UTC on August 1, 2021.

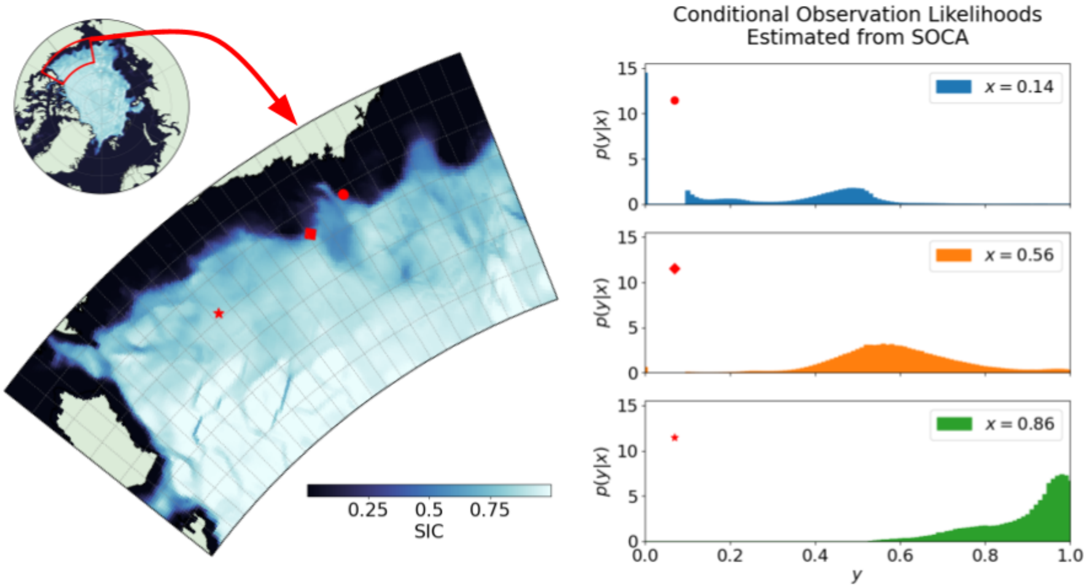

Figure 4: SIC along the marginal ice zone from SOCA experiments valid at 00 UTC August 1, 2021, and estimates of conditional distributions for SIC observations for three different values of the analysis. The red markings indicate locations in the analysis field with the corresponding SIC.

tion errors with no prior knowledge of their distribution. Adopting these functions with a data assimilation scheme that can utilize them fully leads to enhanced analysis performance. Additionally, although our experiments are presented in terms of an unknown observation error, the KECD method does not rely on the presence of a forward operator—as model states that are used for training and conditioning do not need to be represented in terms of the target observations. This is particularly relevant when the relationship between model states and observations is highly nonlinear or when the observation process is known to be a function of multiple state variables. In the case of SIC observations, for example, it may be possible to avoid computing retrievals and instead compute likelihoods for brightness temperatures conditioned on a combination of atmospheric and sea ice model variables. Doing so would eliminate errors associated with the retrieval algorithm and the secondary weather filters. Future research will investigate the merits of the resulting estimated likelihoods versus the ones estimated in this paper by means of Observing System Simulation Experiments (OSSEs) with a more realistic sea ice model.

We have also shown that the diffusion maps component of our approach is sensitive to a number of parameters that can significantly affect the fidelity of the estimated conditional distributions. As of now, we have chosen these parameters experimentally, but we leave for future work the question of how to tune the algorithm in a systematic way. For instance, the bandwidth of the Gaussian kernel used during the diffusion maps algorithm may be determined as a function of the data itself (Berry et al., 2015), and more complex data-driven kernels can be defined which are anisotropic and therefore able to encode information about correlations and differences in scale across the training data (Berry and Sauer, 2016). Such approaches will be particularly important when the KECD method is applied to complex data that encompasses multiple distinct physical variables.

References

Anderson, J. L., 2023: A Quantile-Conserving Ensemble Filter Framework. Part II: Regression of Observation Increments in a Probit and Probability Integral Transformed Space. https://doi.org/10.1175/MWR-D-23-0065.1.

Aronszajn, N., 1950: Theory of Reproducing Kernels. *Transactions of the American Mathematical Society*, **68 (3)**, 337–404, https://doi.org/10.2307/1990404.

Berry, T., D. Giannakis, and J. Harlim, 2015: Nonparametric forecasting of low-dimensional dynamical systems. *Physical Review E*, **91 (3)**, 032 915, https://doi.org/10.1103/PhysRevE.91.032915.

Berry, T., and J. Harlim, 2017: Correcting Biased Observation Model Error in Data Assimilation. https://doi.org/10.1175/MWR-D-16-0428.1.

Berry, T., and T. Sauer, 2016: Local kernels and the geometric structure of data. *Applied and Computational Harmonic Analysis*, **40 (3)**, 439–469, https://doi.org/10.1016/j.acha.2015.03.002.

Burgers, G., P. J. v. Leeuwen, and G. Evensen, 1998: Analysis Scheme in the Ensemble Kalman Filter.

Cavalieri, D. J., P. Gloersen, and W. J. Campbell, 1984: Determination of sea ice parameters with the NIMBUS 7 SMMR. *Journal of Geophysical Research: Atmospheres*, **89 (D4)**, 5355–5369, https://doi.org/10.1029/JD089iD04p05355.

Coifman, R. R., and S. Lafon, 2006: Diffusion maps. *Applied and Computational Harmonic Analysis*, **21 (1)**, 5–30, https://doi.org/10.1016/j.acha.2006.04.006.

Comiso, J. C., 2005: Impacts of the Variability of Ice Types on the Decline of the Arctic Perennial Sea Ice Cover. URL https://ntrs.nasa.gov/citations/20060002674, nTRS Author Affiliations: NASA Goddard Space Flight Center NTRS Meeting Information: Annals of Glaciology, IGS Symposium on Sea Ice; 2005-12-05 to 2005-12-09; undefined NTRS Document ID: 20060002674 NTRS Research Center: Goddard Space Flight Center (GSFC).

Desroziers, G., L. Berre, B. Chapnik, and P. Poli, 2005: Diagnosis of observation, background and analysis-error statistics in observation space. *Quarterly Journal of* *the Royal Meteorological Society*, **131 (613)**, 3385–3396, https://doi.org/10.1256/qj.05.108.

Gloersen, P., and D. J. Cavalieri, 1986: Reduction of weather effects in the calculation of sea ice concentration from microwave radiances. *Journal of Geophysical Research: Oceans*, **91 (C3)**, 3913–3919, https://doi.org/10.1029/JC091iC03p03913.

Hu, C.-C., and P. J. van Leeuwen, 2021: A particle flow filter for high-dimensional system applications. *Quarterly Journal of the Royal Meteorological Society*, **147 (737)**, 2352–2374, https://doi.org/10.1002/qj.4028.

Hu, C.-C., P. J. van Leeuwen, and A. J. Geer, 2024: A non-parametric way to estimate observation errors based on ensemble innovations. *Quarterly Journal of the Royal Meteorological Society*, **150 (761)**, 2296–2315, https://doi.org/10.1002/qj.4710.

Karspeck, A. R., 2016: An Ensemble Approach for the Estimation of Observational Error Illustrated for a Nominal 1° Global Ocean Model. https://doi.org/10.1175/MWR-D-14-00336.1.

Lei, J., and P. Bickel, 2011: A Moment Matching Ensemble Filter for Nonlinear Non-Gaussian Data Assimilation. https://doi.org/10.1175/2011MWR3553.1.

Lorenz, E. N., 1995: Predictability: a problem partly solved. PhD Thesis, ECMWF, Shinfield Park, Reading, publication Title: Seminar on Predictability, 4-8 September 1995 Volume: 1.

Markus, T., and D. J. Cavalieri, 2009: The AMSR-E NT2 Sea Ice Concentration Algorithm : its Basis and Implementation. *Journal of The Remote Sensing Society of Japan*, **29 (1)**, 216–225, https://doi.org/10.11440/rssj.29.216.

Meier, W. N., J. S. Stewart, Y. Liu, J. Key, and J. A. Miller, 2017: Operational Implementation of Sea Ice Concentration Estimates from the AMSR2 Sensor. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, **PP (99)**, https://doi.org/10.1109/JSTARS.2017.2693120.

Meissner, T., and F. J. Wentz, 2012: The Emissivity of the Ocean Surface Between 6 and 90 GHz Over a Large Range of Wind Speeds and Earth Incidence Angles. *IEEE Transactions on Geoscience and Remote Sensing*, **50 (8)**, 3004–3026, https://doi.org/10.1109/TGRS.2011.2179662.

Nyström, E. J., 1930: Über Die Praktische Auflösung von Integralgleichungen mit Anwendungen auf Randwertaufgaben. *Acta Mathematica*, **54 (none)**, 185–204, https://doi.org/10.1007/BF02547521.

Poterjoy, J., 2022: Regularization and tempering for a moment-matching localized particle filter. *Quarterly Journal of the Royal Meteorological Society*, **148 (747)**, 2631–2651, https://doi.org/10.1002/qj.4328.

Poterjoy, J., L. Wicker, and M. Buehner, 2019: Progress toward the Application of a Localized Particle Filter for Numerical Weather Prediction. https://doi.org/10.1175/MWR-D-17-0344.1.

Santer, H., J. Poterjoy, and M. E. Gharamti, 2025: Innovation-Based Methods for Online Estimates of Observation Error Variances within Ensemble Data Assimilation Cycles. https://doi.org/10.1175/MWR-D-24-0242.1.

Snyder, C., T. Bengtsson, P. Bickel, and J. Anderson, 2008: Obstacles to High-Dimensional Particle Filtering. https://doi.org/10.1175/2008MWR2529.1.

Song, L., K. Fukumizu, and A. Gretton, 2013: Kernel Embeddings of Conditional Distributions: A Unified Kernel Framework for Nonparametric Inference in Graphical Models. *IEEE Signal Processing Magazine*, **30 (4)**, 98–111, https://doi.org/10.1109/MSP.2013.2252713.

van Leeuwen, P. J., H. R. Künsch, L. Nerger, R. Potthast, and S. Reich, 2019: Particle filters for high-dimensional geoscience applications: A review. *Quarterly Journal of the Royal Meteorological Society*, **145 (723)**, 2335–2365, https://doi.org/10.1002/qj.3551.

Winitzki, S., 2010: *Linear algebra via exterior products*. Lulu.com, San Bernardino, oCLC: 915400149.

Zappa, C. J., N. J. M. Laxague, S. E. Brumer, and S. P. Anderson, 2019: The Impact of Wind Gusts on the Ocean Thermal Skin Layer. *Geophysical Research Letters*, **46 (20)**, 11 301–11 309, https://doi.org/10.1029/2019GL083687.