A Quantile-Conserving Ensemble Filter Framework. Part III: Data Assimilation for Mixed Distributions with Application to a Low-Order Tracer Advection Model

Jeffrey Anderson[a], Chris Riedel[b], Molly Wieringa[c], Fairuz Ishraque[d], Marlee Smith[a], Helen Kershaw[a]

[a] NCAR/CISL/TDD/DAReS, Boulder, Colorado, jla@ucar.edu

[b] Cooperative Programs for the Advancement of Earth System Science, University Corporation for Atmospheric Research

[c] University of Washington, Department of Atmospheric Sciences, Seattle, Washington

[d] Department of Geosciences, Princeton University, Princeton, New Jersey

Submitted to Monthly Weather Review, November 2023

Abstract

The uncertainty associated with many observed and modeled quantities of interest in Earth system prediction can be represented by mixed probability distributions that are neither discrete nor continuous. For instance, a forecast probability of precipitation can have a finite probability of zero precipitation, consistent with a discrete distribution. However, nonzero values are not discrete and are represented by a continuous distribution; the same is true for rainfall rate. Other examples include snow depth, sea ice concentration, amount of a tracer or the source rate of a tracer. Some Earth system model parameters may also have discrete or mixed distributions. Most ensemble data assimilation methods do not explicitly consider the possibility of mixed distributions. The Quantile Conserving Ensemble Filtering Framework (Anderson 2022, 2023) is extended to explicitly deal with discrete or mixed distributions. An example is given using bounded normal rank histogram probability distributions applied to observing system simulation experiments in a low-order tracer advection model. Analyses of tracer concentration and tracer source are shown to be improved when using the extended methods. A key feature of the resulting ensembles is that there can be ensemble members with duplicate values. An extension of the rank histogram diagnostic method to deal with potential duplicates shows that the ensemble distributions from the extended assimilation methods are more consistent with the truth.

SIGNIFICANCE STATEMENT: Data assimilation is a statistical method that is used to combine information from computer forecasts with measurements of the Earth system. The result is a better estimate of what is occurring in the physical system. As an example, data assimilation is used for making weather predictions. Some Earth system quantities, like precipitation, have special values that can occur very frequently. For instance, zero rainfall is quite common, while any other specific amount of rainfall, say 0.42 inches, is unusual. New data assimilation tools that work well for quantities like this are introduced and should lead to better estimates and predictions of the Earth system.

KEYWORDS: Data assimilation, Ensembles, Uncertainty, Atmospheric Chemistry

1. Introduction

Ensemble data assimilation methods have been widely applied across Earth system applications. The input to the assimilation method is an ensemble of forecasts that is assumed to be a random sample of the uncertainty of a model state vector. Atmospheric data assimilation for numerical weather prediction remains the most common application (Houtekamer and Zhang 2016). In this case, the uncertainty distributions for many variables like temperature, velocity components, and surface pressure are expected to be approximately normal. Many existing ensemble filter algorithms implicitly assume normality (Burgers et al. 1998, Houtekamer and Mitchell 1998, Pham 2001, Anderson 2001) and are very successful for weather prediction applications.

Other types of continuous distributions may be more appropriate for the uncertainty of other variables (Bocquet et al. 2010). For instance, log-normal (Fletcher and Zupanski 2006), gamma and inverse gamma distributions might be more appropriate for variables that are bounded like specific humidity (Bannister et al., 2020). Ensemble filters that can represent gamma and inverse gamma distributions have been developed (Bishop 2016). Other ensemble methods have been developed to transform distributions so that they are more normally distributed (Doron et al. 2013, Kurosawa and Poterjoy 2021), allowing normal ensemble algorithms to work better (Simon and Bertino 2012). The term Gaussian anamorphosis (Bertino et al. 2003) has been applied to some of these methods (Beal et al. 2010, Amezcua and Van Leeuwen 2014). Mixtures of standard continuous distributions like Gaussian kernels (Anderson and Anderson 1999, Grooms 2022) including binormal distributions (Chan et al. 2020) have also been applied.

The uncertainty for some variables is a mixed probability distribution that includes both discrete and continuous parts. As an example, the amount of precipitation that falls during a particular period (Suhaila et al. 2011) might have a discrete probability of being exactly zero in addition to a continuous distribution of being non-zero; the precipitation rate would have a similar mixed distribution. The amount of sea ice, snow cover, chemical tracer, or water in a stream also have mixed distributions along with their source and sink rates. Quantities like the fractional coverage of ice or snow are doubly bounded, and could have a discrete probability of no cover, a discrete probability of complete coverage, and a continuous distribution for all intermediate values. A beta distribution might be appropriate for some doubly bounded quantities.

Anderson (2003) described a two-step algorithm for computing a variety of ensemble Kalman filter algorithms and this methodology was extended for more general problems in Grooms (2022). The input to the first step is an ensemble of estimates of an observed quantity and the likelihood of the observation, while the output is an ensemble of increments due to the observation. The second step is a bivariate algorithm that independently computes increments for each individual model state variable given the increments from step one.

The first part of this quantile conserving ensemble filter framework (QCEFF) paper sequence (Anderson 2022; A22 hereafter) describes the use of quantile conserving ensemble filters for the first step of the two-step algorithm. This allows almost any continuous probability distribution function (PDF) to be used for the computation of observation increments. The second part of the QCEFF sequence (Anderson 2023; A23 hereafter) addresses the second part of the two-step algorithm. It uses a specific variant of anamorphosis, the probit probability integral (PPI) transform (Amezcua and Van Leeuwen 2014), to make the bivariate problem more normal. Again, arbitrary continuous PDF can be used for the probability integral transform portion of the algorithm. Both QCEFF papers include a description of a particular type of nearly non-parametric distribution, the bounded normal rank histogram (BNRH) distribution that can be useful for data assimilation when the details of an appropriate parametric distribution are not known a priori.

A22 provides an example using a discrete distribution that is closely related to the particle filter (Van Leeuwen 2009, Van Leeuwen et al. 2019) and A23 mentions the possibility of using a similar distribution for the PPI transform. However, neither manuscript provides a detailed description of the implementation of the discrete distribution and neither explores mixed distributions. This third part of the QCEFF sequence begins by describing a general framework for using mixed distributions to represent uncertainty in ensemble filters in section 2. When ensemble methods are applied for mixed distributions, ensemble members with identical values for a given state variable are expected to occur. Section 3 extends the results of section 2 to describe a BNRH distribution that works with ensembles with duplicate members. Section 3 also describes an extension of the rank histogram diagnostic tool to ensembles with duplicate members. Section 4 describes an extension of the low-order Lorenz-96 model to include an advected tracer and a source. This model is configured to generate ensembles with duplicate members for both the tracer concentration and source ensemble estimates. Observing system simulation experiments in Section 5 compare the capabilities of several ensemble filter variants in this model. Section 6 provides discussion and conclusions.

2. QCEFF for discrete and mixed probability distributions

The QCEFF developed in A22 for the first part of the two-step ensemble DA algorithm requires finding an appropriate PDF and corresponding cumulative distribution function (CDF) given an ensemble. It requires multiplying the PDF times a likelihood function to get an analysis (posterior) PDF and corresponding CDF. It also requires evaluating CDFs and their inverses; this is also necessary for the probit probability integral (PPI) transforms used for QCEFF implementations of the second part of the two-step algorithm in A23. A22 includes a brief discussion of using a particle filter as the prior generalized PDF and provides an example without carefully defining the algorithm. This section begins by clarifying the application of the QCEFF for discrete probability distributions (like the particle filter), then extends that to mixed probability distributions.

Here, a discrete probability distribution consists of a set of K real numbers, $\{x_i, i = 1, \dots, K\}$ and associated positive real probabilities p_i with

$$\sum_{i=1}^{K} p_i = 1. \tag{1}$$

Suppose a discrete generalized PDF is used as the prior for an observed quantity in data assimilation and the observation likelihood is $L(x)$. The normalizing constant for the product of the prior and the likelihood is

$$S = \sum_{i=1}^{K} L(x_i) p_i. \tag{2}$$

An analysis generalized PDF then has the same $\{x_i\}$ with probabilities

$$p_i^a = L(x_i) p_i / S. \tag{3}$$

To use the QCEFF, it is necessary to evaluate the CDF, and its inverse, corresponding to a discrete generalized PDF. Defining the CDF as the integral from $-\infty$ to x of the generalized PDF leads to discrete jumps at each x_i so that the CDF is not a function. For the QCEFF, a generalized CDF, $\tilde{F}$, that is a function is defined by making the value at x_i the midpoint of the jump,

$$\tilde{F}(x) = \begin{cases} 0 & if\, x < x_1 \\ \sum_{k=1}^{i} p_k \quad if\ x_i < x < x_{i+1}, & i \in \{1, \dots, K-1\} \\ 1 & if\ x > x_K \\ \sum_{k=1}^{i-1} p_k + \frac{p_i}{2} & if\ x = x_i \end{cases} \tag{4}$$

A generalized inverse CDF is defined as

$$\tilde{F}^{-1}(y) = \begin{cases} x_1 & if\ y \le p_1 \\ x_i \quad if \sum_{k=1}^{i-1} p_k < y \le \sum_{k=1}^{i} p_k\,, & i \in \{2, \dots, K\} \end{cases} \tag{5}$$

Note that $x = \tilde{F}^{-1}\left(\tilde{F}(x)\right)$ but $\tilde{F}\left(\tilde{F}^{-1}(y)\right)$ is not necessarily equal to y. With these definitions, it is possible to define a QCEFF that uses any discrete prior, like a particle filter, in observation space for the first part of the two-step filter and for the PPI in the regression step.

As noted in the introduction, mixed distributions are relevant to many geophysical problems. The discrete part of a prior mixed distribution is represented as above except that $\sum p_i = \alpha$; the continuous part of the PDF is $(1-\alpha) f_c(x)$, with $0 < \alpha < 1$. The normalizing constant for the product with a likelihood is

$$S = \alpha \sum_{i=1}^{K} L(x_i)\, p_i + (1-\alpha) \int_{-\infty}^{\infty} L(x) f_c(x) dx \tag{6}$$

The analysis generalized PDF has discrete part as in (3) and the continuous part

$$(1-\alpha) f_c(x) L(x)/S. \tag{7}$$

A generalized CDF corresponding to a mixed PDF is

$$\tilde{F}_m = (1-\alpha) \int_{-\infty}^{x} f_c(x) dx + \alpha \tilde{F}(x) \tag{8}$$

where $\tilde{F}$ is defined in (4). The inverse is clearly defined except at the jumps from the discrete part of the mixed distribution. Define the bounds of the jumps as

$$J_i^- = \begin{cases} (1-\alpha) \int_{-\infty}^{x_1} f_c(x) dx & if\ i = 1 \\ (1-\alpha) \int_{-\infty}^{x_i} f_c(x) dx + \sum_{k=1}^{i-1} \alpha\, p_k, & i \in \{2, \dots, K\} \end{cases} \tag{9}$$

and

$$J_i^+ = J_i^- + \alpha p_i, \quad i \in \{1, \dots, K\} \tag{10}$$

The inverse can be defined between the jump values as

$$\tilde{F}_m^{-1}(y) = x_i \quad for\ J_i^- \le y \le J_i^+ \tag{11}$$

3. **Tools for data assimilation with duplicate ensemble members**

a. Bounded normal rank histogram distribution

The QCEFF described in A22 and A23 requires a CDF to compute observation increments and to do the regression of those increments onto model state variables. The bounded normal rank histogram (BNRH) distribution is an extension of the rank histogram filter distribution for observation space increments (Anderson 2010). A BNRH distribution is particularly useful when the appropriate distribution family is unknown.

A23 describes the PDF, $f(x)$, associated with a BNRH when there are no duplicate ensemble members. An N-member ensemble partitions the real line into N+1 intervals. The interior intervals are bounded on both sides; the intervals on the tails can be bounded on one side only if the quantity itself is not bounded, or bounded on both sides if the quantity is bounded. The BNRH PDF assigns 1/(N+1) probability to each interval. The probability is uniformly distributed over the range of an interior interval. For intervals on the tails, the probability density is part of a normal distribution. The DA algorithms in A22 and A23 require the CDF which is defined in the standard fashion as $F(x) = \int_{-\infty}^{x} f(x)dx$. An example of a BNRH CDF is shown in Figure 1a (reproduced from A23) for a 5-member ensemble.

The definition of the BNRH CDF is extended here for the case when there are ensemble members with duplicate values or when one or more ensemble members have the same value as the upper or lower bound of x. Suppose that possible values of x are bounded below by $B_l \geq -\infty$ and above by $B_u \leq \infty$. Given an N-member ensemble of x with members not necessarily unique, there is at least one ordering of the ensemble values so that $x_i \leq x_{i+1}$ for $i \in \{1, \ldots, N-1\}$. Given such an ordering, define the CDF as:

$$F(x) = \begin{cases} 0 & if x < B_l \\ C(B_l)/[2(N+1)] & if\ x = B_l \\ A_l\Phi(\mu_l, \sigma^2; x) - A_l\Phi(\mu_l, \sigma^2; B_l) & if\ B_l < x < x_1 \\ [i + (x - x_i)/(x_{i+1} - x_i)]/(N+1) & if\ x_i < x < x_{i+1},\ \ i \in \{1, \ldots, N-1\} \\ i/(N+1) + [C(x) - 1]/[2(N+1)] & for\ min\ i\ with\ x = x_i, B_l < x < B_u,\ \ i \in \{1, \ldots, N\} \\ A_u\Phi(\mu_u, \sigma^2; x) - A_u\Phi(\mu_u, \sigma^2; B_u) + 1 & if\ x_N < x < B_u \\ 1 - C(B_u)/[2(N+1)] & if\ x = B_u \\ 1 & if\ x > B_u \end{cases}$$

(12)

$C(x)$ is a function with unbounded real domain and range the whole numbers less than or equal to N, defined as the number of ensemble members with value x. $\Phi(\mu, \sigma^2; x)$ is the CDF of a normal with mean μ and variance σ^2 evaluated at x, and σ^2 is the sample variance of the ensemble. The means and amplitudes of the normal portions are defined as in A23 so that $1/(N+1)$ probability lies between the outermost ensemble member and the bounds. The means are selected so that

$$\Phi(\mu_l, \sigma^2; x_1) = \frac{1}{N+1} \tag{13}$$

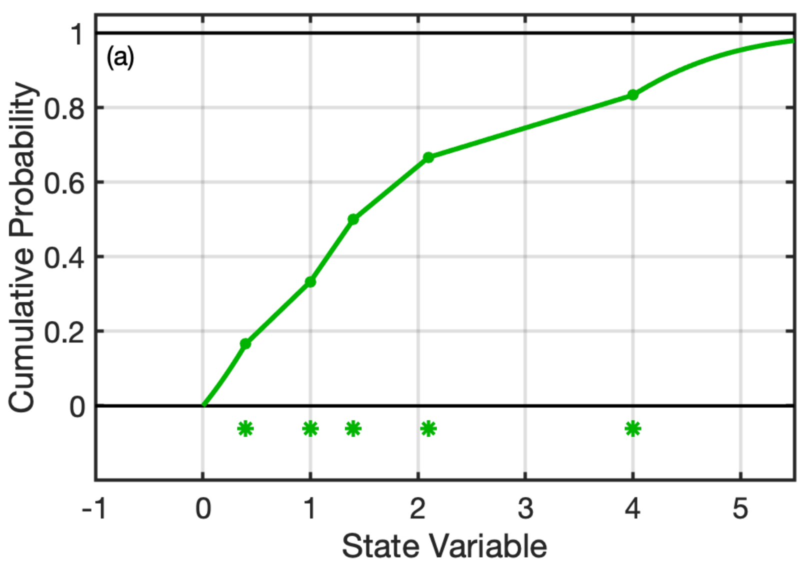
$$\Phi(\mu_u, \sigma^2; x_N) = \frac{N}{N+1} \tag{14}$$

and the amplitudes are

$$A_l = \frac{1}{(N+1)[\Phi(\mu_l, \sigma^2; x_1) - \Phi(\mu_l, \sigma^2; B_l)]} \tag{15}$$

$$A_u = \frac{1}{(N+1)[\Phi(\mu_u, \sigma^2; B_u) - \Phi(\mu_u, \sigma^2; x_N)]} \tag{16}$$

When there are no duplicate ensemble values, $C(x_i) = 1\ \forall x_i$, and no ensemble values equal to the bounds, $C(B_l) = C(B_u) = 0$, the BNRH CDF is equal to the integral from $-\infty$ to x of the BNRH PDF defined in appendix C of A23 and F is invertible. However, where $C(x) > 1$, or if $C(B_l) > 0$ or $C(B_u) > 0$, there is a discrete probability, the derivative $dF(x)/dx$ is undefined, and F is not invertible. It is necessary to define a generalized inverse following the procedure for mixed probability distributions in section 2 (A22 notes the need for a generalized inverse for some other distribution families in which the PDF is 0 over a bounded range of x).



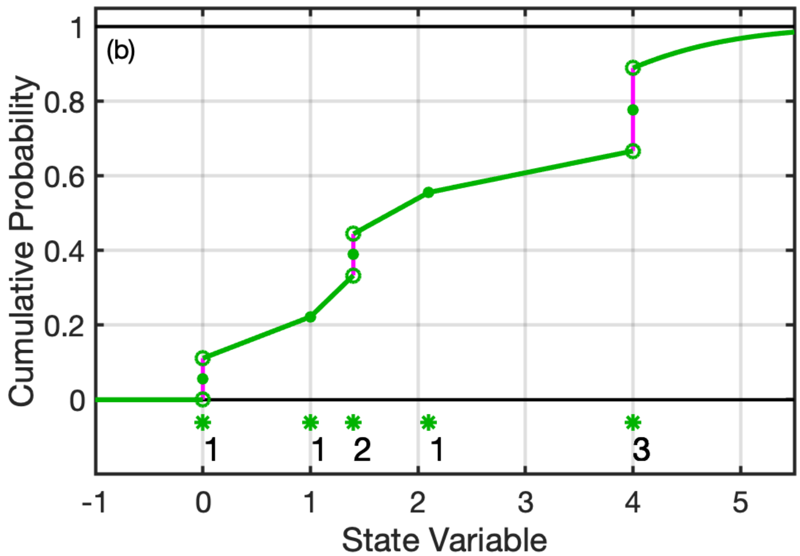


Figure 1: Cumulative distribution functions (green) for a BNRH distribution for a 5-member ensemble (green asterisks) for a variable that is bounded below at zero (a) and for an 8-member ensemble with duplicate values and a member with a value at the bound of zero (b). The number of duplicates is given by the integer next to asterisk. The vertical magenta lines indicate the inverse cumulative distribution function (the quantile function) used for the BNRH. Panel a is reproduced from figure C1a in A23.

An example CDF for an 8-member ensemble with $B_l = x_1 < x_2 < x_3 = x_4 < x_5 < x_6 = x_7 = x_8$ and $B_u = \infty$ is shown in green in Figure 1b. The interval on the upper tail is a portion of a normal CDF. 1/(N+1) probability is uniformly distributed in each interior interval. In non-zero range interior intervals, the CDF is piecewise linear. In the case of the duplicate ensemble members, the range of the interval between them can be thought of as zero and the distribution is discrete. At the point x_3 where there are two ensemble members, there is 1/(N+1) probability while at the point x_6 with three ensemble members, there is 2/(N+1) probability. Generalizing, at any point with D duplicate ensemble members, there is (D-1)/(N+1) discrete probability. Consistent with section 2 and eq. 12, the BNRH CDF at a point with duplicate ensemble members is set to the ‘midpoint’ of the discontinuous jump in the integral of the PDF. For instance, at x_3 the CDF is defined as

$$F(x_3) = \left[\frac{3}{(N+1)} + \frac{4}{N+1}\right]/2. \qquad (17)$$

With this extended definition of the CDF, the quantile computed for ensemble members that share a value is the same. The inverse of the CDF is also needed for the QCEFF algorithms, and it is not uniquely defined with duplicate ensemble members. The method in section 2 leads to defining the inverse as the magenta lines in Fig. 1b.

b. Rank histograms

Consider a sample of $N+1$ numbers composed of an N-member ensemble estimate of a scalar quantity and an additional value, called the verification here. If there are no duplicate values in the sample, the rank of the verification is uniquely defined with an integer value in $\{1, 2, \ldots, N+1\}$. Define a rank weight vector, $W_n,\ \ n = 1, \ldots N+1$ as

$$W_n = \begin{cases} 1 & if\ rank(verification) = n \\ 0 & otherwise \end{cases} \quad (18)$$

Define the sum of the rank weight vector for a collection of M ensembles with verifications as

$$S_n = \sum_{m=1}^{M} W_n^m \quad (19)$$

A histogram of the vector S, commonly called the rank histogram (Anderson 1996, Hamill 2001) is a diagnostic tool for evaluating the consistency of ensemble predictions. If the verification for each ensemble is drawn from the same distribution as the ensemble, the histogram is expected to be statistically uniform. Histograms that are not uniform can provide information about the differences between ensembles and verification. For instance, a U-shaped histogram can indicate under dispersive ensembles (Wilks 2019).

For state variables in many common Earth system DA applications, the probability that the verification duplicates one or more ensemble members is very small, and most discussions of rank histograms have ignored the possibility. However, this is no longer the case for some types of bounded state variables which have mixed probability distributions like the examples discussed in Section 1. If the verification duplicates one or more ensemble members, its rank is no longer uniquely defined by (18). Suppose that D ensemble members have the same value as the verification. When these are removed from the ensemble, the rank of the verification in the $N+1-D$ remaining numbers is uniquely defined, even if there are other duplicate values in the remaining ensemble; let that rank be R. The actual rank in the full ensemble could range from R to $R+D$ since the order of the verification and its duplicates is not uniquely defined. Essentially, there is a $1/(D+1)$ probability that the rank of the verification is any of these values. In this case, define the weight vector as

$$W_n = \begin{cases} 1/(D+1) & if\ R \leq n \leq R+D \\ 0 & otherwise \end{cases} \tag{20}$$

A sum of rank weight vector can be defined for a collection of ensembles as before with (19), and the histogram should be uniform if the verifications are drawn from the same distribution as the ensemble members. Another possible way to define rank histograms for duplicate values is to randomly select one of the ranks between R and R+D and give it the weight of one, however this generates unnecessary random noise compared to the solution in (20).

This treatment of duplicates for rank histograms is essential for application to state variables or true observations in an OSSE like the one in section 4. When rank histograms are used for verifications that are real observed quantities, it is important to account for observational error when generating an appropriate ensemble (Anderson 1996). One way to do this is to add a random sample from an observational error distribution to each ensemble member generated by applying a forward operator to the model state. In many cases, adding in this observation error component would eliminate duplicate values like those that result from bounded state variables in state space. However, if the error distribution is also mixed, duplicates are still expected. Note that a deterministic method similar to (20) can also be developed to account for observational error in the rank histogram.

4. **A tracer advection extension of the Lorenz-96 Model: L96-T**

a. Model description

A low-order model with sensitive dependence on initial conditions, low computational cost, and bounded state variables is useful for testing DA algorithms. The traditional Lorenz-96 model (Lorenz and Emanuel 1998) has been used in many ensemble DA studies including (A22). Here, the Lorenz-96 model is extended to include two additional types of M variables that are collocated with the standard variables, $x_m, m = 0, \ldots, M-1$, on the standard periodic domain. The first type, q_m, represents concentrations of a dimensionless tracer. The second type, s_m, represents a source rate of the tracer with units of tracer amount per time. A function of the standard x variables is treated as a wind field that passively advects the tracer. The velocity at the model grid points at the current time is defined as $v_m = \bar{V} + \tilde{V} x_m$ where $\bar{V}$ is a specified constant mean velocity, $\tilde{V}$ is a specified multiplying constant that controls the average magnitude of wind perturbations, and $\tilde{V} x_m$ is an anomalous velocity at gridpoint m. Velocities are expressed with units of nondimensional distance per nondimensional time. A nondimensional location is assigned to each grid point in the model so that the distance between neighboring grid points is 1 (note that this is different from many previous Lorenz-96 studies where the distance between grid points is defined as 1/M).

The time evolution of the standard variables, x_m, is identical to that used in the basic Lorenz-96 model (Lorenz and Emanuel 1996). The time evolution of the nonnegative tracer concentration used here is:

$$q_m^+ = max[(q_m^{adv} + s_m \Delta t)e^{-E\Delta t} - C\Delta t, 0] \qquad (21)$$

where q_m^+ is the tracer concentration at grid point m at the next time step, q_m^{adv} is the advected concentration, s_m is the source rate at grid point m at the current time, E is an exponential damping time, C is a constant sink rate, and Δt is the timestep.

The advection of tracer is computed using an upstream semi-Lagrangian method. The computation of q_m^{adv}, the advected concentration at the next time at grid point m given the wind field at the current time, v_m, and the concentrations at the current time, q_m, proceeds as follows:

1. A preliminary upstream target location is defined as $T = m - v_m \Delta t$,
2. The fractional location of the target between the bounding grid points is $p = T - \lfloor T \rfloor$ where the brackets indicate the floor,

3. The indices of the grid points bounding the target location are computed as $L = mod(\lfloor T \rfloor, M)$ and $U = mod(L+1, M)$,
4. The advected concentration is $q_m^{adv} = (1-p)q_L + pq_U$

The specified source is a function of grid point and model time with units of amount per time. For experiments here, there is a time constant source with rate 5 at grid point 1 and all other grid points have zero source at all times

$$s_m = \begin{cases} 5 & if\ m = 1 \\ 0 & otherwise \end{cases}.$$

b. L96-T example

All results here use the standard 4th order Runge-Kutta time stepping algorithm, the nondimensional $\Delta t = 0.05$ with an associated dimensional time step of 3600s as done in many previous studies, and $M = 40$ grid points. The L96 forcing parameter $F = 8$. The mean velocity $\bar{V} = 0$ and the velocity perturbation multiplier $\tilde{V} = 5$, while the constant sink $C = 0.1$, and the exponential sink $E = 0.25$.

Figure 2 shows a time series of the wind field, v_m, as a function of the model grid point; since $\bar{V} = 0$, this is just $\tilde{V} = 5$ times the standard L96 state variables, x_m. The well-known group and phase velocity of the L96 model can be seen.

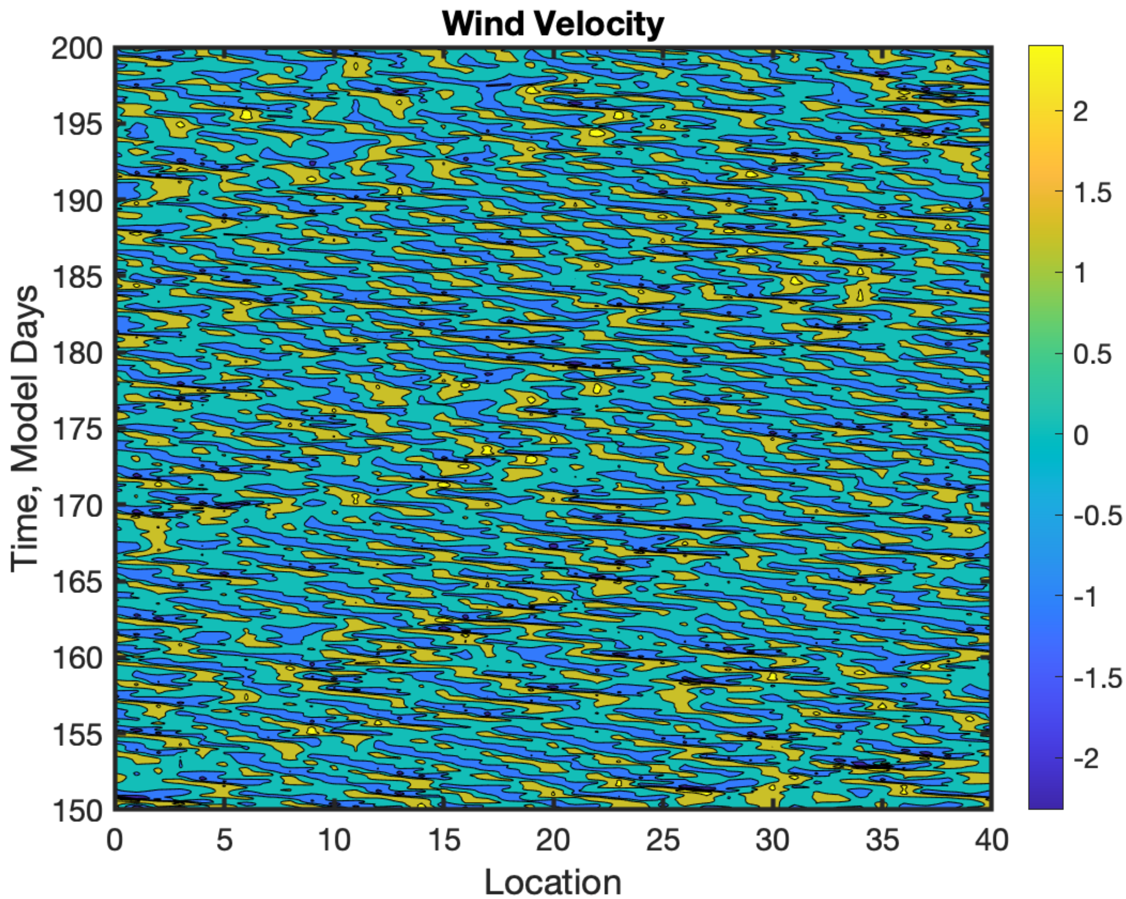


Figure 2: Wind velocity from the L96 Tracer Advection model for times between day 150 and 200 in the truth run as a function of model grid point. Units are distance hr $^{-1}$.

Figure 3 shows the tracer concentration corresponding to the wind field. Plumes of tracer are advected away from the source at grid point 1. The velocity is positive more often than negative, so plumes extend more frequently and further to the right. However, the wind field is sometimes negative leading to shorter plumes extending to the left. The white areas in the plot have zero tracer concentration, so a mixed distribution is appropriate. It is rare for plumes to extend clear across the domain with this only happening twice in the figure. This behavior is roughly analogous to what one might see with a point source in the midlatitudes. It is possible to get a variety of other behaviors for the tracer by changing the model parameter values.

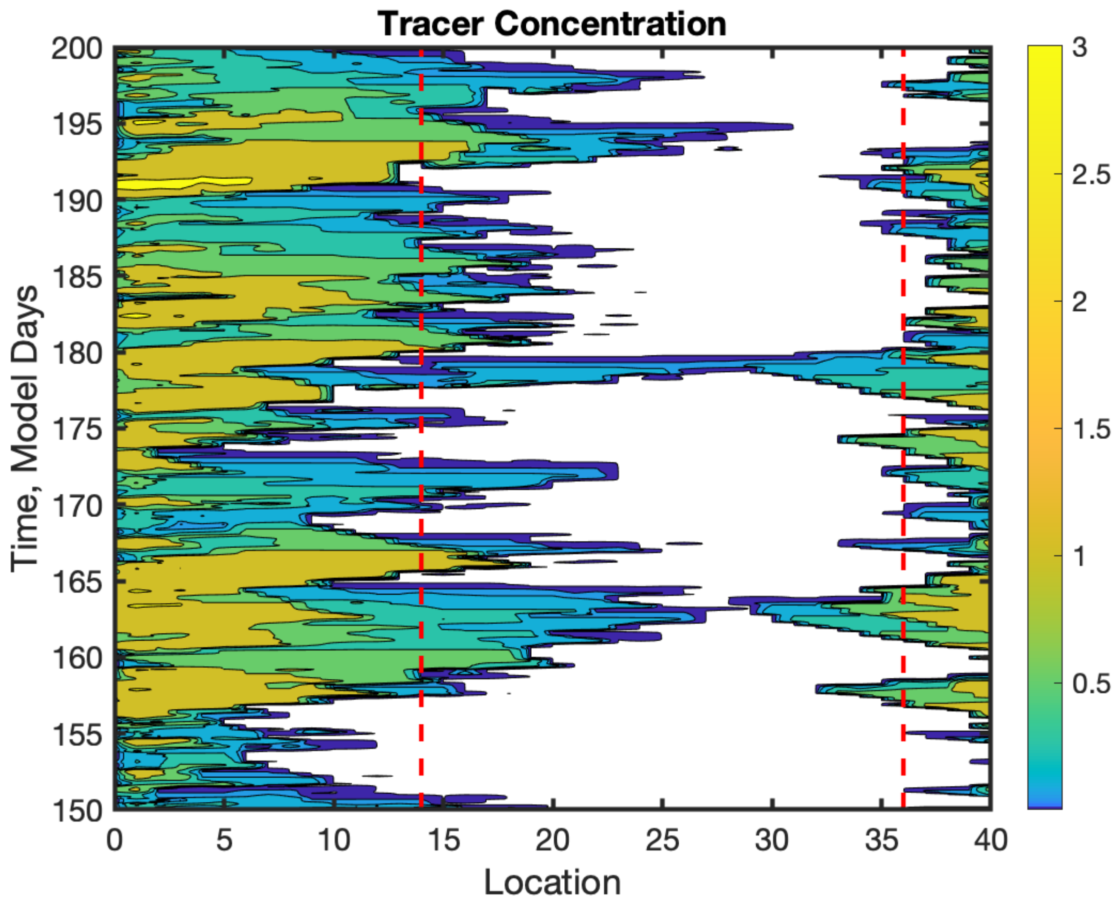


Fig. 3: As in figure 2 but for the tracer concentration (nondimensional). The red dashed lines highlight grid points with additional diagnostics presented in figures 4, 5 and 6. White areas have zero concentration.

5. Data assimilation experiments

The model integration described in the previous section is used as the truth run for a series of observing simulation system experiments (OSSEs). The L96-T model is first integrated for 16500 hrs (5500 3-hour advances) starting from a default initial state to generate a tuning initial state. The default initial state has $x_1 = 1$ and all other x state variables are 0; all concentration variables are 0. The model is integrated for an additional 16500 hrs from the tuning initial state generating synthetic observations every 3 hrs. Forty randomly located observing sites are selected for the L96 standard state, and a different set of 40 randomly located sites for tracer concentration observations (see Figs 7d and 8d). Observations are taken by linearly interpolating to the site location from the two nearest grid points. For the standard state observations error is simulated by adding a random draw from a normal distribution with mean 0 and variance 10. For tracer observations error is simulated by adding a random draw from a truncated normal distribution with variance 0.1 and lower bound of 0 (A23, appendix D).

Three different observing networks are explored: assimilating only standard state observations, assimilating only tracer concentration observations, and assimilating both standard and tracer observations. Two different model configurations are evaluated. In the first, every ensemble member has the true value of the tracer source variables. In the second, the tracer source variables are unknown, and every ensemble member has its own (not necessarily unique) time evolving estimate.

All assimilation experiments use the adaptive inflation algorithm of Gharamti (2018) with an inflation damping of 0.9. All experiments also use a Gaspari Cohn localization with the same constant halfwidth for all observations. Seven halfwidth tuning assimilation experiments are done for each case, where a case is defined by the observing network, whether the source is known or unknown, and the ensemble size (20, 40, 80 or 160). As in A23, the halfwidths tested are $\{0.075, 0.1, 0.125, 0.175, 0.2, 0.4, \infty\}$. These tuning assimilations start from the tuning initial condition and assimilate for 5500 3-hour intervals. Initial ensembles for the standard state variable are generated by adding a random draw from a normal distribution with mean 0 and standard deviation 0.01 to the truth value for each variable. Initial ensemble members for the tracer variables are all equal to the truth. For the case with known sources, all ensemble members for the source variables are equal to the truth. For the case with unknown sources, ensemble members for the source are set to a random draw from a normal distribution with mean 2.5 and standard deviation 2.5; if the draw is less than 0 the source is set to 0 so that the resulting ensembles are generally mixed distributions with several members that are 0. Results from the first 500 assimilation steps are discarded and the prior ensemble mean, time mean RMSE is computed for the standard state and tracer variables for the remaining 5000 steps. For the state only observing network, the localization that minimizes the state RMSE is selected; for the other observing networks, the localization that minimizes the tracer RMSE is selected.

The model truth is then integrated for an additional 16500 hours from the end of the tuning integration with synthetic observations generated in the same way. Initial conditions for ensembles are also generated in the same way as for the tuning experiments. Assimilation experiments are performed for each case using the tuned localization and assimilating every 3 hours. The first 500 steps are discarded, and results are available for the final 5000 assimilation steps. The spread for all quantities appears to be spun up after fewer than 100 assimilation steps for all experiments.

Four different assimilation algorithms are applied to each case using the QCEFF. As noted in A23, a complete description of a QCEFF assimilation algorithm requires information about the first step where increments are computed for observed variables and the second step where those increments are regressed onto state variables. The QCEFF uses a probit probability integral transform (PPI) before doing the regression (A23). Table 1 specifies the details of the four algorithms which are referred to as an EAKF, RHF, PQBNRH, and DUAL. Note that the normal likelihood used for the q variable in the EAKF is a normal with the same variance as the truncated normal observation error distribution for q. As noted in A23, using a normal for the PPI transform is equivalent to no transform. The BNRH CDFs all have a lower bound of 0 and no upper bound, consistent with the nature of the tracer concentration and source variables.

	EAKF	RHF	PQBNRH	DUAL
x obs. CDF	Normal	RH	RH	Normal
x likelihood	Normal	Normal	Normal	Normal
x PPI CDF	None	None	RH	None
q obs. CDF	Normal	BNRH	BNRH	BNRH
q likelihood	Normal	Truncated Normal	Truncated Normal	Truncated Normal
q PPI CDF	None	None	BNRH	BNRH

s PPI CDF	None	None	BNRH	BNRH

Table 1: Assimilation settings for each of the four algorithms explored. For the x and q variables, the continuous CDF and form of the likelihood used for computing observation space increments are listed with RH referring to a rank histogram distribution without bounds and BNRH referring to a bounded normal rank histogram distribution with a lower bound at zero. The continuous distribution used as part of the PPI transform used when regressing observation increments onto state variable increments is also listed.

a. Known source results

Unless otherwise noted, all results shown are for analysis, rather than forecast, variables. Also, results shown are for the network observing both standard state and tracer observations unless otherwise noted. Figure 4 shows a time series from the EAKF and PQBNRH algorithm 80-member assimilations for tracer at grid point 14, which is highlighted by a red dashed line in Fig. 3. The EAKF ensemble in Fig. 4a represents all the plumes that occur, but also represents two plumes between days 150 and 160 that are not real. The ensemble is strongly biased towards larger values at some times, in particular around days 168, 173, and 178. The PQBNRH results in Fig. 4b also capture all real plumes with smaller values for the two false plumes, but do not have the strongly biased periods.

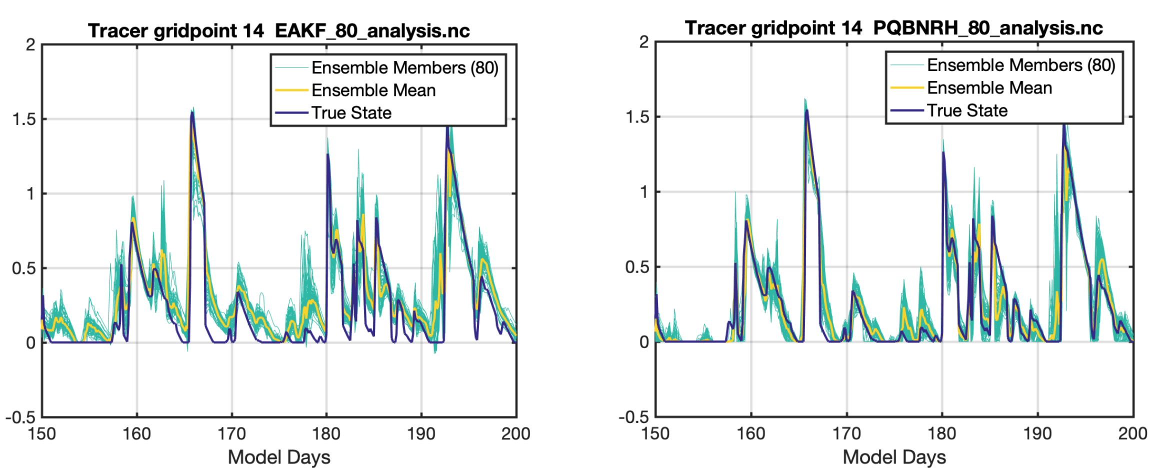


Fig. 4: Time series of the tracer at grid point 14. Dark curve is the truth and is the same in both panels. The dark green curves are the 80 analysis ensemble members, and their mean is in yellow, for an EAKF (left) and a PQBNRH (right); the tracer is nondimensional.

Figure 5 shows rank histograms over all 5000 assimilation steps for concentration at grid point 14. The EAKF and RHF algorithms result in very strongly biased histograms with the truth very often less than the smallest ensemble member. The results for the PQBNRH and DUAL are radically different. Both have histograms that are nearly uniform except for the two outermost bins. The PQBNRH has more cases where the truth is larger than the largest ensemble member while the DUAL algorithm has more cases where the truth is smaller than the smallest member; however, it is difficult to evaluate whether these differences are statistically significant.

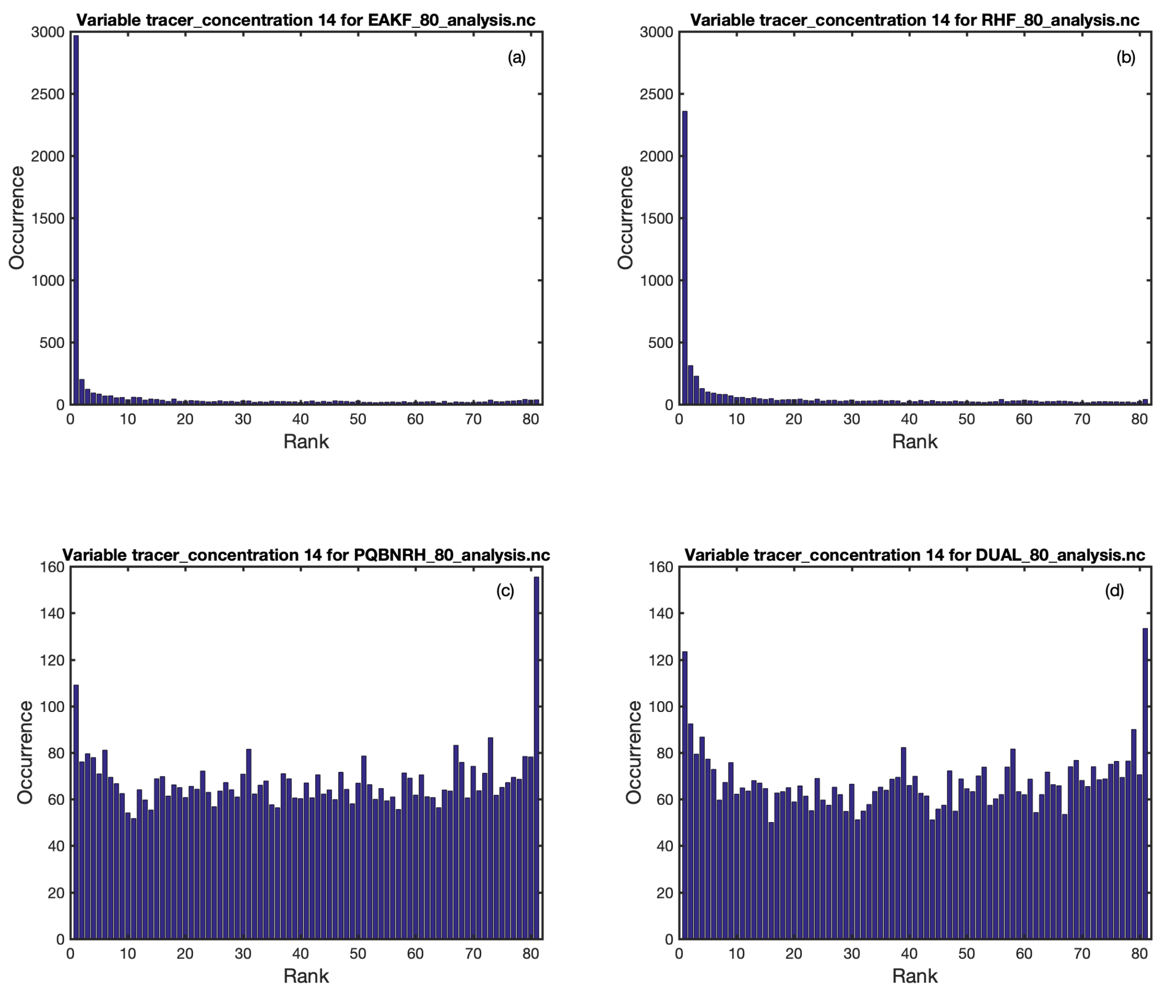


Fig. 5: Rank histograms for 80-member analysis concentration at grid point 14 for an EAKF (a), RHF (b), PQBNRH (c) and a DUAL filter with an EAKF for the wind and a BNRH for the concentration (d). Note the different vertical axes in the top and bottom rows.

Fig. 6 shows time series of the EAKF and PQBNRH assimilation results for grid point 36 which is also highlighted in Fig. 3. At this grid point, plumes are less frequent, primarily arriving from the right. There are extended periods when the true concentration is 0. The EAKF represents all true plumes, however, there are several instances where the ensemble is strongly biased towards larger concentration than the truth, and several times when negative ensemble members occur; this cannot happen with the PQBNRH. The EAKF never has any ensemble members that are exactly zero and never has duplicate ensemble members. The PQBNRH also captures all real plumes and has fewer instances of false plumes. The PQBNRH has several periods when many ensemble members are exactly 0 and some periods where all members are zero, all at times when the truth is also zero. Results for the RHF are similar to those for the EAKF, and results for DUAL are similar to those for the PQBNRH in figures 4 and 6 so these are not displayed.

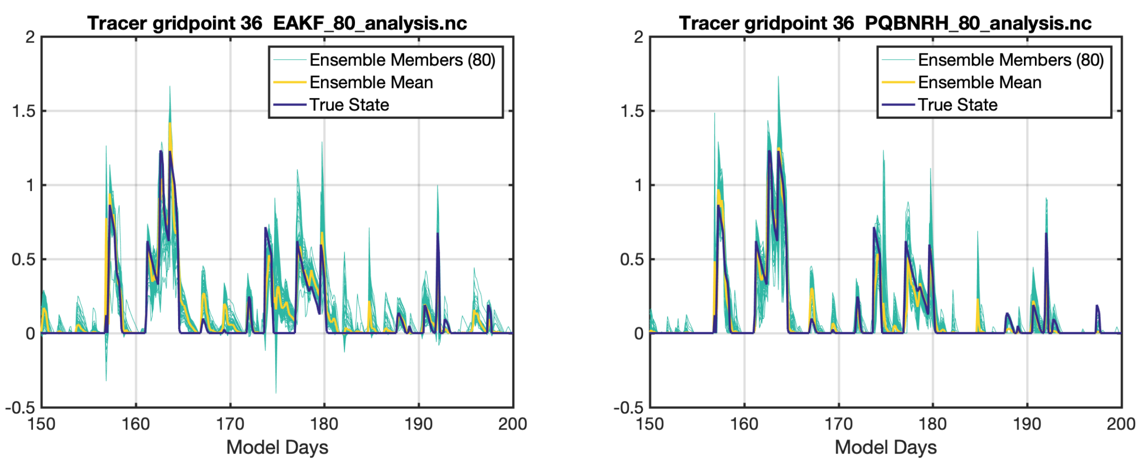


Fig. 6: As in figure 4 but for grid point 36.

Fig. 7 shows summary results for ensemble mean tracer RMSE over all 5000 assimilation steps for the four algorithms and four ensemble sizes studied. In general, the results for the PQBNRH and DUAL algorithms are statistically indistinguishable. The same is true for the EAKF and RHF algorithms. However, in general the PQBNRH/DUAL algorithms have lower RMSE. The RMSE is largest to the left of the source at grid point 0 where the true concentration is most variable, and smaller far from the source where concentration is smaller. There are not large differences as a function of ensemble size; larger ensembles generally have only slightly smaller RMSE. It is unclear why ensemble size is not more important here.

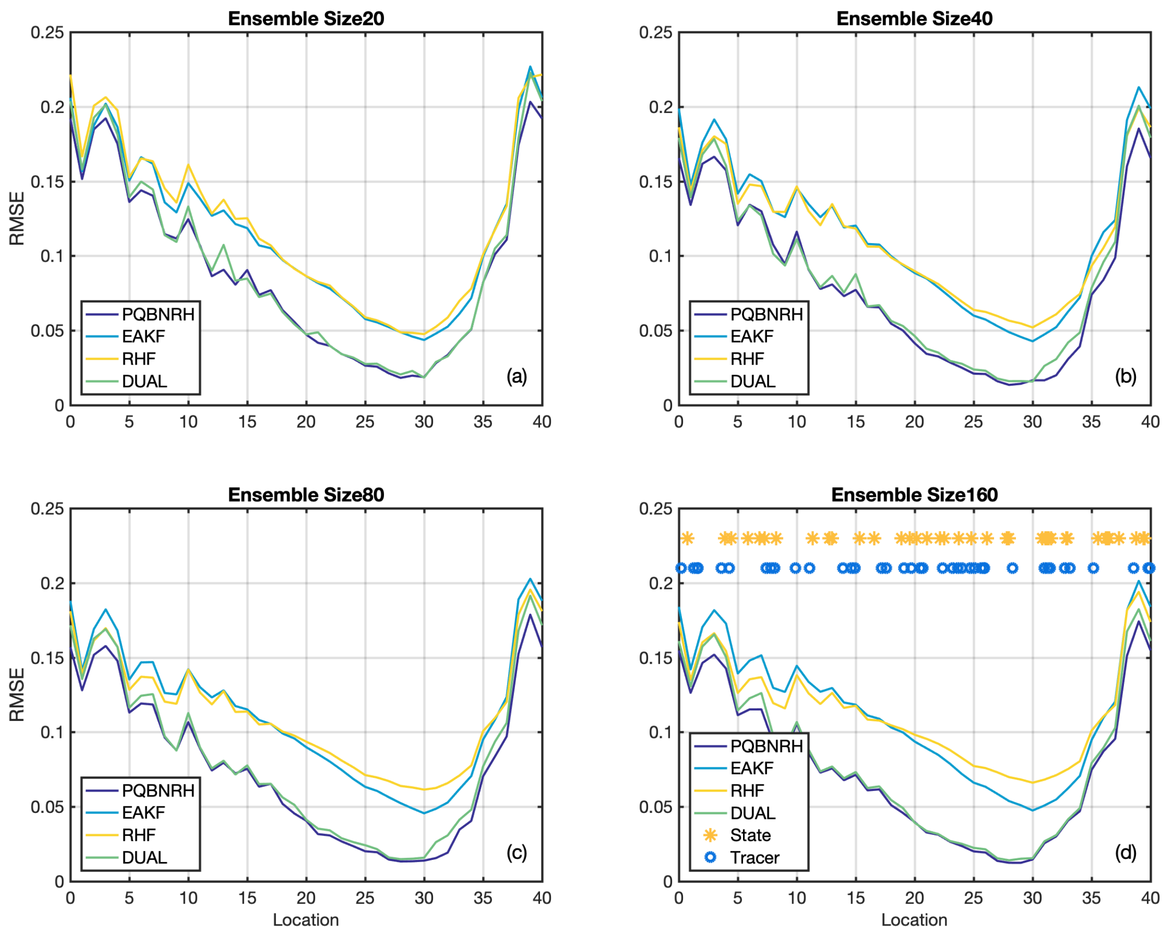
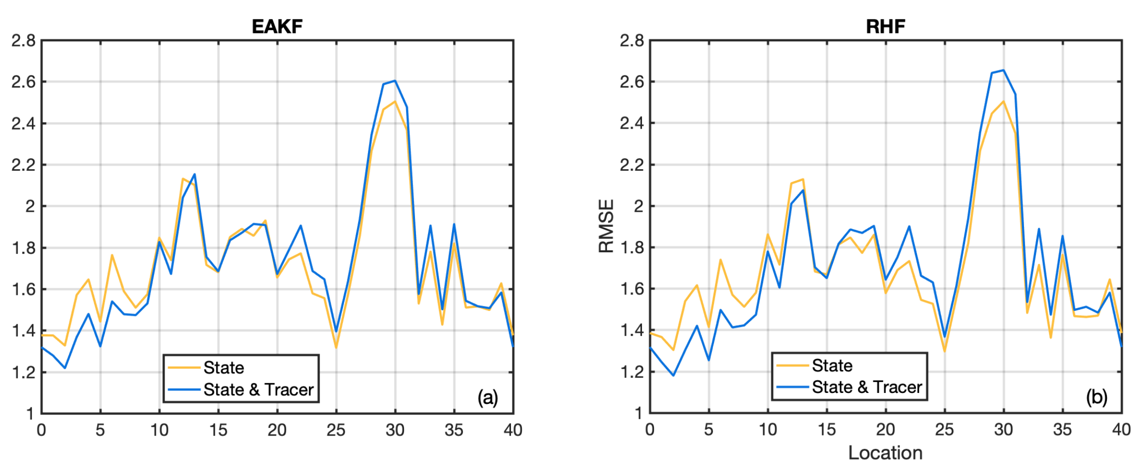


Fig. 7: Ensemble mean, time mean RMSE as a function of grid point for the analysis tracer concentration for four filter algorithms for ensemble size 20 (a), 40 (b), 80 (c) and 160 (d). The locations of the 40 observing stations are shown in (d) for state (yellow circles) and tracer concentration (blue asterisks).

It is obvious that assimilating standard state observations that improve the estimate of the winds will result in improved estimates of the tracer concentrations. However, the impact of tracer observations on the standard state variables is less clear. Assimilations for the network observing only tracer produced tracer analysis estimates that have much larger RMSE than those just discussed, although smaller than the RMSE from an unconstrained control ensemble run. The tracer only network resulted in standard variable RMSE that was only slightly smaller than the RMSE from an unconstrained control.

A comparison of the standard variable RMSE from the observing network with only standard state observations to the network with both standard and tracer observations is shown in Fig. 8 for the four algorithms. The RMSE for the standard observation only network has larger RMSE near grid point 30 and smaller RMSE near grid points 25 and 1. This is due to the random observing site locations (Fig. 8d). The RMSE is smaller for the PQBNRH than for any of the other algorithms; note that the EAKF and DUAL are identical for the standard observation network.

When tracer observations are added in, all four algorithms produce reduced RMSE for the left part of the domain. The EAKF and RHF produce increased RMSE in the right part of the domain. The PQBNRH and DUAL produce roughly equivalent RMSE in the right part of the domain. In the left part of the domain, plumes with large spatial and temporal gradients occur near the source. These provide information about the flow field that is advecting the plume and lead to the reduced RMSE for the standard state. Because there is often very little or no tracer in the right part of the domain, observations of the tracer are expected to provide very little additional information. The increase in error in the EAKF and RHF suggests that deficiencies in these algorithms cause the use of these low information observations to degrade the ensemble estimate.



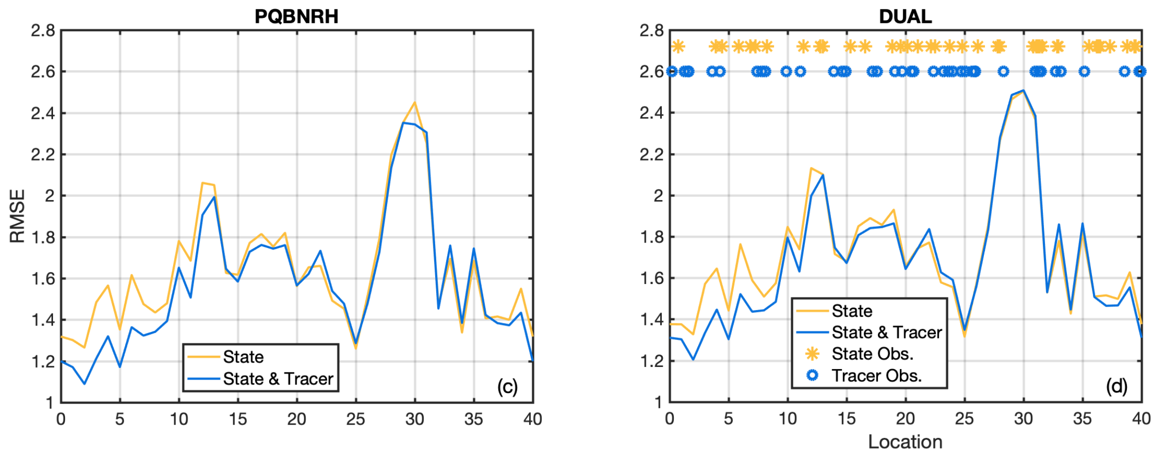


Fig 8: Ensemble mean, time mean RMSE as a function of grid point for the standard L96 state for experiments that assimilate only observations of the standard state, and experiments that also assimilate the tracer concentration, shown for an EAKF (a), RHF (b), PQBNRH (c), and a DUAL filter with an EAKF for the wind and a BNRH for the concentration (d). The locations of the 40 observing stations are shown in (d) for state (yellow asterisks) and tracer concentration (blue circles).

b. Unknown source results

In these experiments, the source is not known and is estimated by the assimilation algorithms. Results are only discussed for the network observing both standard state and tracer observations. There is no time tendency model for the tracer. The prior ensembles can have their spread increased by the adaptive inflation. Nevertheless, in all experiments, the spread becomes increasingly small for the source at all grid points. The source variables are only impacted by concentration observations since the source and the state field should not be meaningfully correlated.

Figure 9 shows the natural logarithm of the absolute value of the error for each ensemble member and the ensemble mean error at the grid point with the nonzero source in the truth for the EAKF and the PQBNRH. Both reduce the ensemble mean error, but the reduction is much larger for the PQBNRH. Because of the collapse of spread, both algorithms eventually have strongly biased estimates and would produce corresponding rank histograms.

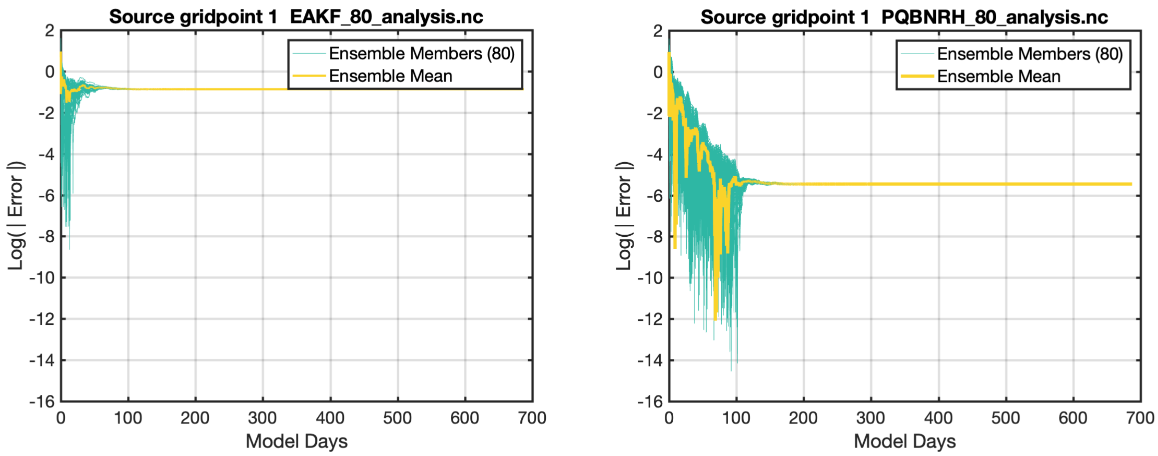


Fig. 9: Spatial mean of the natural logarithm of the absolute value of the error of the ensemble mean (yellow) and each of the 80 ensemble members (green) as a function of time for the source at grid point 1 which has a true value of 5 (units hr^{-1}) for the EAKF (left) and the PQBNRH (right).

Figure 10 shows the evolution of the RMSE for grid point 21 which has zero true source. The RMSE for the EAKF is smaller than it was for grid point 1. The error for the PQBNRH decreases throughout the 5000 assimilation steps. As the assimilation continues, more and more ensemble members have values of exactly zero; eventually all ensemble members are zero and the error of the ensemble mean, and all individual ensembles is zero. At both grid points 1 and 21, the RMSE for the standard state and concentration variables for the PQBNRH are nearly identical to those for the known source experiments since the since the source is so accurately estimated. Results are somewhat degraded for the EAKF which has larger errors in its source estimates.

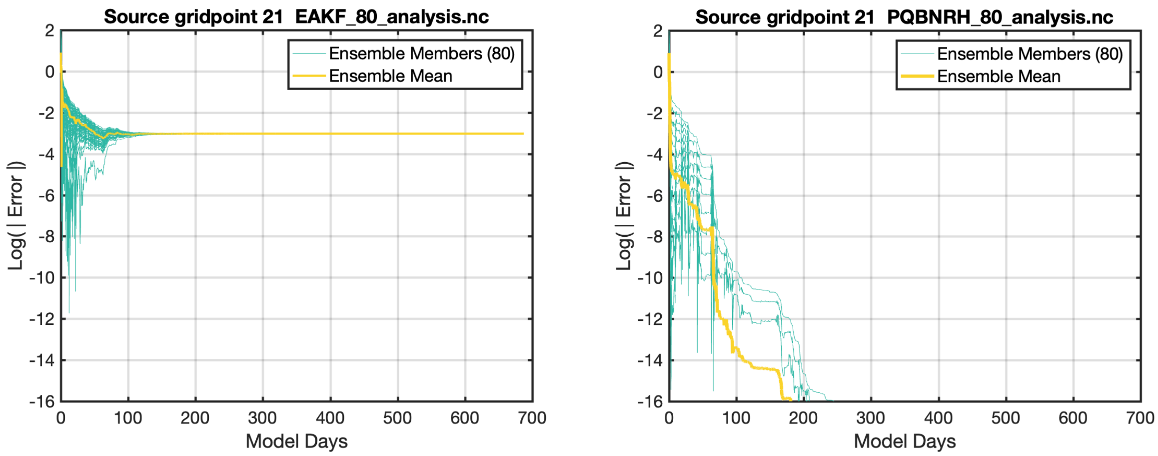


Fig. 10: As in 9 for grid point 21 which has zero true source.

Figure 11 shows the RMSE for the source as a function of grid point for each of the four algorithms and all four ensemble sizes. The EAKF and RHF produce roughly comparable results that have a small dependence on ensemble size. The errors do not have a strong dependence on the grid point. The PQBNRH and DUAL are also very similar but have more dependence on both ensemble size and grid point. The smallest errors occur for grid points close to the non-zero source at grid point 1. The RMSE actually increases with ensemble size in these areas. This is due to the rate at which ensemble members become exactly zero which appears to be similar across ensembles so that the fraction of nonzero members at a given time increases with ensemble size. Larger errors are found for the source point itself and for points far from the source. The estimate at point 1 varies little with ensemble size. The RMSE for points remote from the source gets smaller and less noisy with increasing ensemble size.

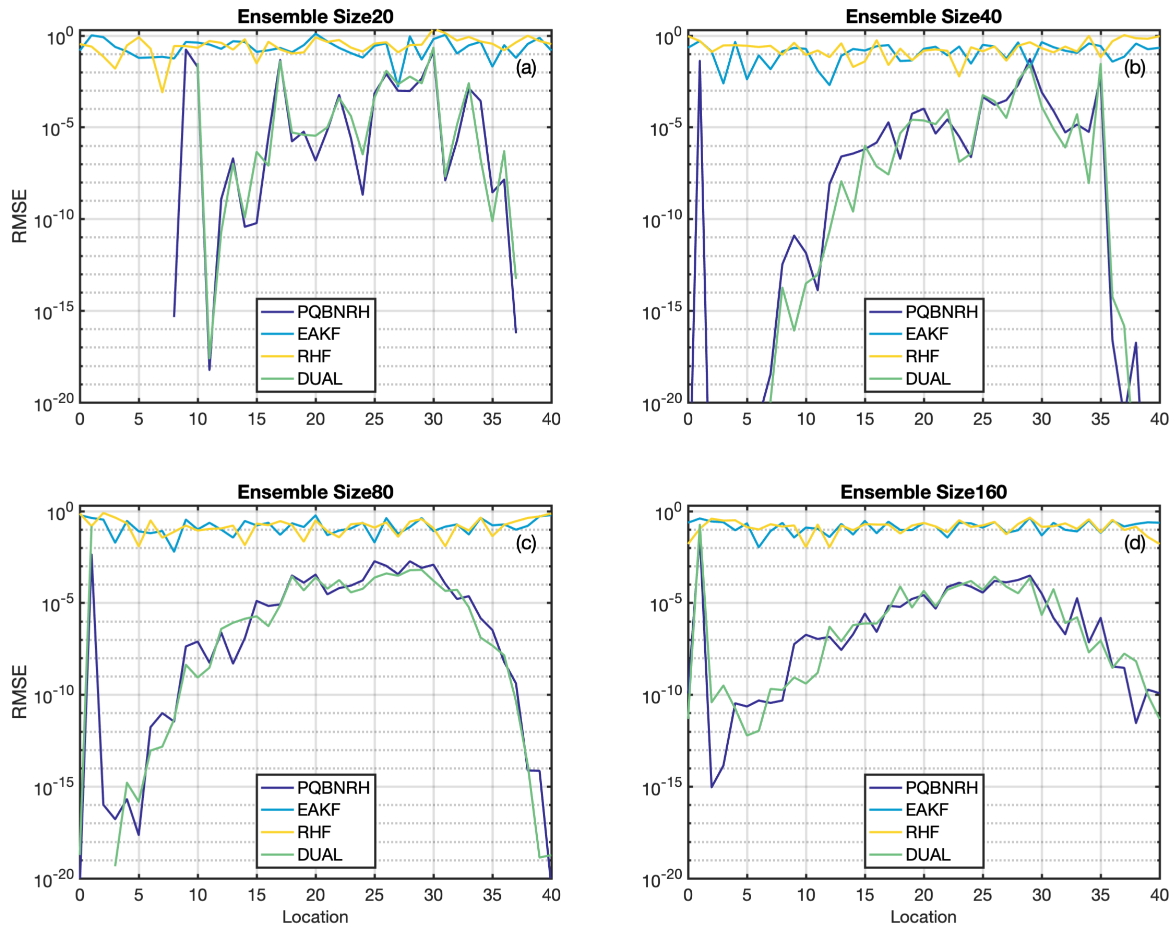


Fig. 11: Ensemble mean time mean RMSE as function of grid point for tracer source for four algorithms for ensemble size 20 (a), 40 (b), 80 (c) and 160 (d). Values that are not plotted for the PQBNRH and DUAL algorithms are less than 10^{-20} including many that are exactly zero.

6. Discussion and conclusions

The QCEFF has been extended to deal with model and observed variables with mixed probability distributions. This capability is especially relevant for bounded quantities like precipitation (Lien et al. 2013), tracer concentrations and sources, and areal coverage (Wieringa et al., 2023 in press; Riedel and Anderson 2023 in press). It may also be useful for estimating model parameters with data assimilation (Gharamti et al. 2016); the tracer source in the L96-T version used here is essentially equivalent to a model parameter.

The nearly non-parametric BNRH distribution has also been extended to handle duplicate ensemble members that are expected to occur for variables with mixed distributions. The rank histogram diagnostic tool was also extended to deal with duplicate ensemble members. An extension of the Lorenz-96 low-order dynamical system that includes an idealized advective tracer with local sources was developed to test the new algorithms. This L96-T model should also provide challenging tests for other data assimilation algorithms including variational methods and particle filters.

Results show that the BNRH works better than the EAKF or RHF for an OSSE with the L96-T model. The RMSE is smaller for the bounded tracer concentration and source variables when they are close to the bounds as might be expected. Results are also better when these variables are not close to the bounds and for the unbounded standard state variables from L96. The RH and PQBNRH algorithms use the BNRH distribution to compute observation increments. However, the RH uses standard linear regression when updating state variables while the PQBNRH includes the PPI transform using the BNRH distribution for the probability integral transform. The RH results are similar to the EAKF results in this case, while the PQBNRH is better for all variables and locations demonstrating that the PPI is a crucial part of the improved performance. The DUAL case uses an EAKF for the L96 state which has no bounds and is expected to be approximately normal. There is some indication that the PQBNRH is slightly better than the DUAL algorithm, but differences are not quantitatively significant. This suggests a strategy of using the BNRH distribution for bounded variables but a normal distribution for other variables may be useful for large model applications.

The BNRH as described allows duplicate ensemble members and the data assimilation process can create additional duplicates; this happened for both concentration and source variable ensembles in the OSSEs here. However, the assimilation process cannot eliminate duplicates. It can change the value of ensemble members that are exactly at a bound in the prior ensemble. This means that the model must eliminate duplicates if appropriate. That happens in experiments here and is most clearly seen in figure 6b where all ensemble members are zero at some times but not at subsequent times. Further investigation into methods that would allow the assimilation to eliminate duplicates is warranted but would require making a priori assumptions about the expected errors associated with a given ensemble size.

All the OSSEs here were performed using the Data Assimilation Research Testbed (DART: Anderson et al. 2009) which implements the QCEFF including the BNRH; the parallel algorithm of Anderson and Collins (2007) was used. The results here only examined the use of normal or BNRH distributions. DART software can support arbitrary distributions and currently supports gamma, inverse gamma, log-normal, beta, and particle filter distributions. Previous work on assimilation of bounded quantities has proposed using distributions like the log-normal, gamma, and inverse gamma. However, the L96-T OSSE explored here presents specific challenges for using these other distributions. The log-normal and inverse gamma distributions do not have any probability at zero. This is clearly inappropriate for the mixed distributions in the OSSE where much of the probability can be at 0 at some times. The gamma distribution can have probability at zero. However, if the likelihood is a gamma distribution, the corresponding observation error distribution is inverse gamma (Bishop 2016, A22). This means that observations of the bounded quantities would not be able to have any probability at zero. This is clearly problematic for the bounded quantities with realistic instruments. Further work on explicitly using mixed distributions, for instance a combination of a log-normal with a discrete distribution, for applications like this is beyond the scope of this report.

The computational cost of the QCEFF algorithms including the BNRH is discussed in detail in A23. There is almost no additional cost associated with allowing duplicate ensemble members so the A23 analysis still applies. As noted there, the additional cost of a BNRH compared to an EAKF can be significant, especially for low-order model applications. As discussed in Anderson (2019) and A23, much of this cost is associated with the need to sort the ensemble members for each state variable. However, the sorting order often changes little between assimilation steps. Caching the sort order and then using sorts that are efficient for nearly sorted data can potentially result in large computational cost reductions, but these methods have not yet been implemented in DART.

The low-order model results here suggest that there may be significant improvements when the BNRH is used for bounded quantities in large Earth system applications. Initial tests in sea ice (Wieringa et al. 2023 in press) and chemical transport models will be investigated in subsequent work. Other types of nearly non-parametric distributions, for instance various kernels (Grooms 2022, Anderson and Anderson 1999) can also be developed in DART and should be compared to the BNRH results.

Acknowledgements. This material is based upon work supported by the National Center for Atmospheric Research, which is a major facility sponsored by the National Science Foundation under Cooperative Agreement 1852977. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the author and do not necessarily reflect the views of the National Science Foundation. Thanks to Ian Grooms, Joseph Chan, Hristo Chipilski, Ben Gaubert and the DAReS team for helpful discussions about this material.

Data availability statement

The Lorenz-96 results were generated with DART code that can be found at: https://github.com/NCAR/DART/releases/tag/MWR_QCEFF_Part3.

REFERENCES

Amezcua, J. and P. J. Van Leeuwen, 2014: Gaussian anamorphosis in the analysis step of the EnKF: a joint state-variable/observation approach. *Tellus A: Dynamic Meteorology and Oceanography*, 66. DOI: 10.3402/tellusa.v66.23493

Anderson, J. L., 1996: A Method for Producing and Evaluating Probabilistic Forecasts from Ensemble Model Integrations. *J. Climate*, **9**, 1518–1530, https://doi.org/10.1175/1520-0442(1996)009<1518:AMFPAE>2.0.CO;2.

Anderson, J. L., 2001: An ensemble adjustment Kalman filter for data assimilation. *Mon. Wea. Rev.*, **129**, 2884-2903.

Anderson, J. L., 2003: A local least squares framework for ensemble filtering. *Mon. Wea. Rev.*, **131**, 634-642.

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

Anderson, J. L., 2019: A Nonlinear Rank Regression Method for Ensemble Kalman Filter Data Assimilation. *Mon. Wea. Rev.*, **147**, 2847–2860, https://doi.org/10.1175/MWR-D-18-0448.1.

Anderson, J. L., 2022: A Quantile-Conserving Ensemble Filter Framework. Part I: Updating an Observed Variable. *Mon. Wea. Rev.*, **150**, 1061–1074, https://doi.org/10.1175/MWR-D-21-0229.1.

Anderson, J. L., 2023: A Quantile-Conserving Ensemble Filter Framework. Part 2: Updating an Observed Variable. *Mon. Wea. Rev.*, **151**, 2759–2777, https://doi.org/10.1175/MWR-D-23-0065.1

Anderson J. L., and S. L. Anderson, 1999: A Monte Carlo implementation of the nonlinear filtering problem to produce ensemble assimilations and forecasts. *Mon. Wea. Rev.*, **127**, 2741–2758.

Anderson, J. L., and N. Collins, 2007: Scalable implementations of ensemble filter algorithms for data assimilation. *J. Atmos. Oceanic Technol.*, **24**, 1452–1463.

Anderson, J., T. Hoar, K. Raeder, H. Liu, N. Collins, R. Torn and A. Arellano, 2009: The Data Assimilation Research Testbed. *Bul. Amer. Met. Soc.*, **90**, 1283-1296.

Bannister, R. N., H. G. Chipilski, and O. Martinez-Alvarado, O., 2020. Techniques and challenges in the assimilation of atmospheric water observations for numerical weather prediction towards convective scales. *Q J R Meteorol Soc*., **146**. https://doi-org.cuucar.idm.oclc.org/10.1002/qj.3652

Beal, D., P. Brasseur, J. M. Brankart, Y. Ourmieres, and J. Verron, 2010: Characterization of mixing errors in a coupled physical biogeochemical model of the North Atlantic:

Implications for nonlinear estimation using Gaussian anamorphosis. *Ocean Sci*., **6**, 247–262.

Bertino, L., G. Evensen and H. Wackernagel, 2003: Sequential Data Assimilation Techniques in Oceanography. *International Statistical Review*, **71**, 223-241.

Bishop, C. H., 2016: The GIGG-EnKF: Ensemble Kalman filtering for highly skewed non-negative uncertainty distributions. *Q. J. R. Meteorol. Soc*., **142**, 1395-1412. doi:10.1002/qj.2742

Bocquet, M., C. A. Pires, and L. Wu, 2010: Beyond Gaussian Statistical Modeling in Geophysical Data Assimilation. *Mon. Wea. Rev.*, **138**, 2997-3023. https://doi.org/10.1175/2010MWR3164.1.

Burgers, G., P. J. van Leeuwen, and G. Evensen, 1998: Analysis scheme in the ensemble Kalman filter. *Mon. Wea. Rev*., **126**, 1719-1724.

Chan, M., J. L. Anderson, and X. Chen, 2020: An Efficient Bi-Gaussian Ensemble Kalman Filter for Satellite Infrared Radiance Data Assimilation. *Mon. Wea. Rev.*, **148**, 5087–5104, https://doi.org/10.1175/MWR-D-20-0142.1.

Doron, M., P. Brasseur, J. M. Brankart, S. N. Losa, and A. Melet, 2013: Stochastic estimation of biogeochemical parameters from Globcolour ocean colour satellite data in a North Atlantic 3D ocean coupled physical–biogeochemical model. *J. Marine Systems*, **117**, 81–95.

Fletcher, S., and M. Zupanski, 2006: A data assimilation method for log-normally distributed observational errors. *Quart. J. Roy. Met. Soc*., **132**, 2505 - 2519. 10.1256/qj.05.222.

Gharamti, M., 2018: Enhanced Adaptive Inflation Algorithm for Ensemble Filters. *Mon. Wea. Rev.*, **146**, 623–640, https://doi.org/10.1175/MWR-D-17-0187.1.

Gharamti, M. E., A. Samuelsen, L. Bertino, E. Simon, A. Korosov, and U. Daewel, 2016: Online tuning of ocean biogeochemical model parameters using ensemble estimation techniques: Application to a one-dimensional model in the North Atlantic. Journal of Marine Systems, 168, 1-16. https://doi.org/10.1016/j.jmarsys.2016.12.003.

Grooms, I., 2022: A comparison of nonlinear extensions to the ensemble Kalman filter. *Comput Geosci* **26**, 633–650. https://doi.org/10.1007/s10596-022-10141-x

Hamill, T. M., 2001: Interpretation of Rank Histograms for Verifying Ensemble Forecasts. *Mon. Wea. Rev.*, **129**, 550–560, https://doi.org/10.1175/1520-0493(2001)129<0550:IORHFV>2.0.CO;2.

Houtekamer, P. L., and H. L. Mitchell, 1998: Data assimilation using an ensemble Kalman filter technique. *Mon. Wea. Rev.*, **126**, 796-811.

Houtekamer, P. L., and F. Zhang, 2016: Review of the ensemble Kalman filter for atmospheric data assimilation, *Mon. Wea. Rev.*, ***144***, 4489-4532.

Kurosawa, K., and J. Poterjoy, 2021: Data Assimilation Challenges Posed by Nonlinear Operators: A Comparative Study of Ensemble and Variational Filters and Smoothers. *Mon. Wea. Rev.*, **149**, 2369–2389, https://doi.org/10.1175/MWR-D-20-0368.1.

Lien, G.Y., E. Kalnay, and T. Miyoshi, 2013: Effective assimilation of global precipitation: Simulation experiments. Tellus A 65, DOI: 10.3402/tellusa.v65i0.19915

Lorenz, E. N. and K. A. Emanuel, 1998: Optimal sites for supplementary weather observations: Simulation with a small model. *J. Atmos. Sci.*, **55**, 399-414.

Pham, D.T., 2001: Stochastic methods for sequential data assimilation in strongly nonlinear systems. *Mon. Wea. Rev.,* **129**, 1194–1207. Doi:10.1175/1520-0493

Riedel, C., 2023: In press

Simon, E. , and L. Bertino, 2012: Gaussian anamorphosis extension of the DEnKF for combined state parameter estimation: application to a 1D ocean ecosystem model. J. Marine Syst., **89**, 1–18.

Suhaila, J., K. Ching-Yee, Y. Fadhilah and F. Hui-Mean, 2011:Introducing the Mixed Distribution in Fitting Rainfall Data. *Open Journal of Modern Hydrology*, **1**, 11-22. doi: 10.4236/ojmh.2011.12002.

Van Leeuwen, P. J., 2009: Particle filtering in geophysical systems. *Mon. Wea. Rev.*, **137**, 4089-4114. doi.org/10.1175/2009MWR2835.1

Van Leeuwen, P. J., H. R. Künsch, L. Nerger, R. Potthast, and S. Reich, 2019: Particle filters for high-dimensional geoscience applications: A review. *Quart. J. Roy. Met. Soc.*, **149**, 2335-2365. doi.org/10.1002/qj.3551

Wilks, D. S., 2019: Indices of Rank Histogram Flatness and Their Sampling Properties. *Mon. Wea. Rev.*, **147**, 763–769, https://doi.org/10.1175/MWR-D-18-0369.1.