Improving Ensemble Data Assimilation through Probit-space Ensemble Size Expansion for Gaussian Copulas (PESE-GC)

Man-Yau Chan[1,2]

[1]Department of Geography, The Ohio State University, Columbus, Ohio, 43210, USA
[2]Advanced Study Program, National Center for Atmospheric Research, Boulder, Colorado, USA

Correspondence: Man-Yau Chan (chan.1063@osu.edu)

Abstract. Small forecast ensemble sizes (< 100) are common in the ensemble data assimilation (EnsDA) component of geophysical forecast systems, thus limiting the error-constraining power of EnsDA. This study proposes an efficient and embarrassingly parallel method to generate additional ensemble members: the Probit-space Ensemble Size Expansion for Gaussian Copulas (PESE-GC; "peace gee see"). Such members are called "virtual members". PESE-GC utilizes the users' knowledge of the marginal distributions of forecast model variables. Virtual members can be generated from any (potentially non-Gaussian) multivariate forecast distribution that has a Gaussian copula. PESE-GC's impact on EnsDA is evaluated using the 40-variable Lorenz 1996 model, several EnsDA algorithms, several observation operators, a range of EnsDA cycling intervals and a range of forecast ensemble sizes. Significant improvements to EnsDA ($p < 0.01$) are observed when either 1) the forecast ensemble size is small (≤ 20 members), 2) the user selects marginal distributions that improves the forecast model variable statistics, and/or 3) the rank histogram filter is used with non-parametric priors in high forecast spread situations. These results motivate development and testing of PESE-GC for EnsDA with high-order geophysical models.

1 Introduction

Geophysical forecast models are often computationally expensive to run. As a result, geophysical ensemble data assimilation (EnsDA) typically uses <100 forecast ensemble members. Such small forecast ensemble sizes result in sampling errors that degrade the performance of EnsDA. As such, low-cost methods that introduce additional ensemble members (henceforth, "virtual members") to the original forecast members (henceforth, "forecast members") have the potential to improve EnsDA.

Several types of ensemble expansion methods have been proposed in the literature, all of which have strengths and weaknesses. The first type is random draws from climatology (Castruccio et al., 2020; El Gharamti, 2020; Lei et al., 2021). Though computationally efficient, this type of ensemble expansion method cannot generate members with flow-dependent ensemble statistics.

An alternative type of ensemble expansion method is to aggregate forecast ensemble members across time (Xu et al., 2008; Yuan et al., 2009; Huang and Wang, 2018; Gasperoni et al., 2022). Though this type of method efficiently produces members with flow-dependent statistics, the number of virtual members created is limited (Huang and Wang, 2018; Gasperoni et al., 2022).

A third type of ensemble expansion method is to search a historical catalog for forecast states similar to the current forecast or observations (Van den Dool, 1994; Tippett and Delsole, 2013; Monache et al., 2013; Wan and Van Der Merwe, 2000; Grooms, 2021; Sun et al., 2022). The virtual members resulting from this search have flow-dependent statistics. Though such methods are historically expensive to employ, ongoing research may render them affordable in the near future (e.g., Sun et al. (2024)).

Ensemble modulation (Bishop and Hodyss, 2009, 2011; Bishop et al., 2017; Kotsuki and Bishop, 2022; Wang et al., 2021) is the fourth type of ensemble expansion method. Such methods expand ensembles by combining a localization matrix with the original ensembles' perturbations (see Bishop and Hodyss (2009) for details). Though the expanded ensemble possesses the same mean and variance as the original ensemble, the expanded ensemble's kurtosis can be much larger than the original ensemble's (see the Supplement). In other words, the expanded ensemble's kurtosis is likely biased. If nonlinear observation operators are applied on the expanded ensemble, this kurtosis bias will result in biased expanded ensemble observation statistics (personal communication with Craig Bishop and Lili Lei).

The shortcomings of existing ensemble expansion methods motivate the development of a new ensemble expansion method. This study proposes an new ensemble expansion method that explicitly utilizes the users' knowledge of prior marginals: the Probit-space Ensemble Size Expansion for Gaussian Copulas (PESE-GC). PESE-GC constructs virtual members using a generalization of the efficient and scalable Gaussian resampling algorithm of Chan et al. (2020) (henceforth, the CAC2020 algorithm). Unlike existing methods, PESE-GC efficiently and scalably generates an unlimited number of virtual members with flow-dependent statistics. Furthermore, the PESE-GC produces virtual members that are consistent with user-specified prior marginals, and handles multivariate Gaussian distributions and many multivariate non-Gaussian distributions. PESE-GC is applied before running any observation operators for EnsDA, and the analyzed virtual members are discarded before generating the next forecast ensemble (see Fig. 1). To the author's knowledge, no other ensemble expansion methods simultaneously possesses the same efficiency, scalability, generality and flow-dependency as PESE-GC.

The remainder of this publication is divided into five sections. Section 2 discusses the formulation of PESE-GC, and its computational complexity and scalability. PESE-GC's impacts on EnsDA are discussed and illustrated in Section 3. PESE-GC is then tested with EnsDA using the Lorenz 1996 model in Section 4. Section 5.1 discusses an important caveat of the PESE-GC method. This publication then ends in Section 6 with a summary and a discussion of avenues for future research.

2 Formulation of PESE-GC

This section begins with reviewing the CAC2020 algorithm. Then, the CAC2020 algorithm is generalized to handle arbitrary piecewise-continuous marginal distributions (i.e., 1D distributions) using probit probability integral (PPI) transforms. Finally, the computational complexity and scalability of PESE-GC is discussed.
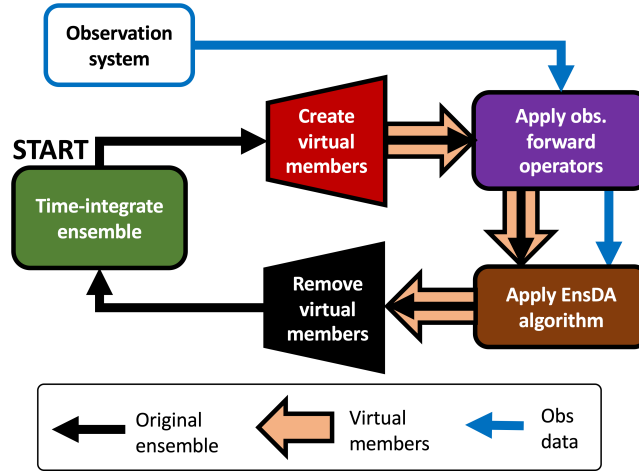
Figure 1. An illustration of how PESE-GC can be integrated into a typical EnsDA cyclic workflow. This workflow is meant to be read starting from the green box labeled "START". The arrows indicate the movement of various kinds of information (see legend). For example, the fat orange arrows indicate that the virtual members are created by PESE-GC (red polygon), passed to the observation operators (purple rounded box), passed to the EnsDA algorithm (brown rounded box), and then removed before applying the forecast model (black polygon). "Obs" stands for "observation" and "ens." stands for "ensemble".

2.1 The CAC2020 algorithm

55

The CAC2020 algorithm constructs Gaussian-distributed virtual members through linear combinations of the forecast ensemble perturbations. The resulting expanded ensemble has the same mean state and covariance matrix as the forecast ensemble. The CAC2020 algorithm is first formulated by Chan et al. (2020) and a more comprehensive derivation is presented in Chapter 7 of Chan (2022).

60 To write down the CAC2020 algorithm, a notation similar to Ide et al. (1997) is used. Suppose $\boldsymbol{x}$ is an N_x-dimensional column vector representing a forecast model state, $\left\{\boldsymbol{x_1^f}, \boldsymbol{x_2^f}, \ldots, \boldsymbol{x_{N_E}^f}\right\}$ represents an ensemble of N_E forecast model states,

$$\boldsymbol{x_n^{f'}} \equiv \boldsymbol{x_n^f} - \frac{1}{N_E} \sum_{n'=1}^{N_E} \boldsymbol{x_{n'}^f}, \ \forall n = 1, 2, \ldots, N_E, \tag{1}$$

N_V virtual members $\left\{\boldsymbol{x_1^v}, \boldsymbol{x_2^v}, \ldots, \boldsymbol{x_{N_V}^v}\right\}$ are constructed, and

$$\boldsymbol{x_m^{v'}} \equiv \boldsymbol{x_m^v} - \frac{1}{N_V} \sum_{m'=1}^{N_V} \boldsymbol{x_{m'}^v}, \ \forall m = 1, 2, \ldots, N_V. \tag{2}$$

65 Note that the mean of the virtual members is the same as the mean of the forecast members. In other words,

$$\frac{1}{N_V} \sum_{m=1}^{N_V} \boldsymbol{x_m^v} = \frac{1}{N_E} \sum_{n=1}^{N_E} \boldsymbol{x_n^f}. \tag{3}$$

2.1.1 Step 1 of the CAC2020 algorithm

The CAC2020 algorithm constructs N_V virtual members from N_E ensemble members using a three-step procedure. First, an $N_E \times N_V$ matrix of linear combination coefficients ($\boldsymbol{E}$) is generated by evaluating

$$\boldsymbol{E} \equiv \gamma \mathbf{1}_{N_E \times N_V} + \mathrm{Chol}\left(\boldsymbol{C_E}\right)\left\{\mathrm{Chol}\left(\boldsymbol{WW}^\top\right)\right\}^{-1}\boldsymbol{W}. \tag{4}$$

Here,

$$\gamma \equiv \frac{1}{N_V}\left(\sqrt{\frac{N_E + N_V - 1}{N_E - 1}} - 1\right), \tag{5}$$

$\mathbf{1}_{N_E \times N_V}$ is an $N_E \times N_V$ matrix of ones, $\mathrm{Chol}\left(\cdot\right)$ represents Cholesky decomposition (which produces a lower-triangular matrix), $\boldsymbol{C_E}$ is an $N_E \times N_E$ matrix defined by

$$\boldsymbol{C_E} \equiv \frac{N_V}{N_E - 1}\boldsymbol{I_{N_E}} - \gamma^2 N_V \mathbf{1}_{N_E \times N_E}, \tag{6}$$

$\boldsymbol{I_{N_E}}$ is the $N_E \times N_E$ identity matrix, and $\mathbf{1}_{N_E \times N_E}$ is an $N_E \times N_E$ matrix where every element is one. Finally, $\boldsymbol{W}$ is an $N_E \times N_V$ matrix whose (i,j)-th element is defined by

$$W_{i,j} = \omega_{i,j} - \frac{1}{N_V}\sum_{\ell=1}^{N_V}\omega_{i,\ell} \tag{7}$$

where all $\omega_{i,j}$s are identically and independently distributed (i.i.d.) samples drawn from the standard normal distribution. In other words, all $\omega_{i,j}$s are mutually independent.

2.1.2 Step 2 of the CAC2020 algorithm

The CAC2020 algorithm's second step is to generate $\left\{\boldsymbol{x}_1^{v'}, \boldsymbol{x}_2^{v'}, \ldots, \boldsymbol{x}_{N_V}^{v'}\right\}$ by evaluating

$$\begin{bmatrix}\boldsymbol{x}_1^{v'} & \cdots & \boldsymbol{x}_{N_V}^{v'}\end{bmatrix} = \begin{bmatrix}\boldsymbol{x}_1^{f'} & \cdots & \boldsymbol{x}_{N_E}^{f'}\end{bmatrix}\boldsymbol{E}. \tag{8}$$

2.1.3 Step 3 of the CAC2020 algorithm

In the third and final step, the CAC2020 algorithm generates $\left\{\boldsymbol{x}_1^v, \boldsymbol{x}_2^v, \ldots, \boldsymbol{x}_{N_V}^v\right\}$ by evaluating

$$\boldsymbol{x}_m^v \equiv \boldsymbol{x}_m^{v'} + \frac{1}{N_E}\sum_{n'=1}^{N_E}\boldsymbol{x}_{n'}^f, \quad \forall m = 1, 2, \ldots, N_V. \tag{9}$$

2.1.4 On $\boldsymbol{W}$ used in the CAC2020 algorithm's step 1

Note that this study's (and the CAC2020's) $\boldsymbol{W}$ differs from that of Chan et al. (2023) (henceforth, "CAC2023"). This is because the CAC2023's $\boldsymbol{W}$ [defined in Eq. (6) of CAC2023] generates virtual members with undesirable non-Gaussian properties. The CAC2023's $\boldsymbol{W}$ can be written as

$$\text{CAC2023's } \boldsymbol{W}: \quad W_{ij} \equiv \delta_{i,j} - \frac{1}{N_V} \tag{10}$$
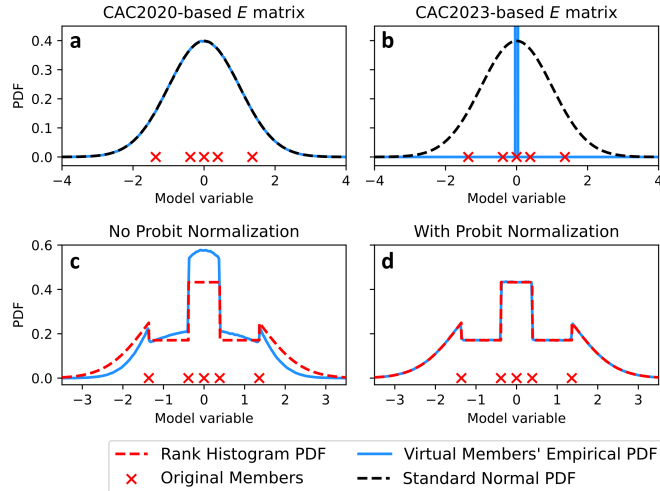
Figure 2. Plots demonstrating the impacts of various heuristic choices in the formulation of PESE-GC. Panels a and b respectively demonstrate drawing virtual members using CAC2020's W [Eq. (7)] and CAC2023's W [Eq. (10)]. The former produces Gaussian virtual members while the latter produces non-Gaussian virtual memebrs. Panels c and d demonstrate the importance of ensuring that the forecast ensemble in probit-space has unit variance. If the probit-space forecast ensemble's variance is not unity, the virtual members generated by PESE-GC will deviate from the fitted marginal PDFs (rank histogram in the case of panels c and d).

where $\delta_{i,j}$ is the Kronecker-delta.

To illustrate the issue with the CAC2023's W, suppose 5 forecast members are drawn from a standard normal distribution and 10^7 virtual members are generated using the CAC2020 algorithm with the CAC2023's W. Though the expanded ensem-
95 ble's mean and variance are correct (zero and unity respectively), the virtual members' histogram-estimated PDF (blue curve in Fig. 2b) is incorrect (not standard normal). In contrast, using the W defined in Eq. (7) results in virtual members that follow the standard normal PDF (Fig. 2a). As such, this study uses the CAC2020's W instead of the CAC2023's W.

2.1.5 Properties of the CAC2020 algorithm

The CAC2020 algorithm is efficient and scales well with parallel computing. This is because steps 2 and 3 of the CAC2020
100 algorithm (Sections 2.1.2 and 2.1.3) are embarrassingly parallel and have computational complexities that scale linearly with N_x. If E is generated offline (i.e., not part of the EnsDA procedure) and then read into the EnsDA program, the CAC2020 algorithm is reduced to only evaluating steps 2 and 3.

The CAC2020 algorithm also produces expanded ensembles with same ensemble means and ensemble covariances as the forecast ensembles. In other words, the rank of the expanded ensemble's covariance matrix is the same as that of the forecast
105 ensemble. Future work can explore ways to incorporate localization into the expanded ensemble's covariance matrix.

Furthermore, the CAC2020 algorithm always generates Gaussian-distributed virtual members: even if the actual forecast distribution is highly non-Gaussian, the virtual members' distribution will still be Gaussian. The CAC2020 algorithm thus

degrades the ensemble statistics in situations where the forecast distribution is non-Gaussian. This degradation limits the usefulness of the CAC2020 algorithm for situations with non-Gaussian forecast distributions.

110   Note that, except for the mean and covariance, the expanded ensemble's central moments (i.e., higher-order moments; e.g., skewness) likely differ from the forecast ensemble's. More specifically, the expanded ensemble's central moments will be closer to those of Gaussian distributions (e.g., zero skewness) than the forecast ensemble's central moments. This is because the virtual members are effectively drawn from a Gaussian distribution. If the forecast distribution is indeed a Gaussian distribution, then the expanded ensemble likely has better moments than the forecast ensemble.

2.2   The PESE-GC procedure

The CAC2020 algorithm is limited to generating Gaussian-distributed virtual members. PESE-GC overcomes this limitation by combining probit probability integral (PPI) transforms and their inverses with the CAC2020 algorithm. A PPI transform transforms any univariate distribution with a continuous CDF into a standard normal distribution, and the inverse PPI transform reverses the process. The quantity resulting from applying the PPI transform on a random variable is called "probit" and

120   the coordinate space occupied by probits is called "probit-space". Such transforms are often used in Gaussian anamorphosis (Amezcua and Van Leeuwen, 2014; Grooms, 2022).

To define the PPI transform, suppose $F_i(x_i)$ represents the CDF of the i-th model variable x_i ($i = 1, 2, \ldots, N_x$), $\Phi^{-1}(q)$ represents the inverse CDF of the standard normal (q represents any quantile), and ϕ_i represents the i-th model probit. The double appearance of index i in $F_i(x_i)$ is deliberate – the CDF varies with the chosen model variable. Note that $\Phi^{-1}(q)$ is

125   sometimes called "the probit function" or the "quantile function of the standard normal". The conversion from x_i to ϕ_i (i.e., the PPI transform) is

$$\phi_i \equiv \Phi^{-1}\left(\, F_i(x_i)\, \right). \tag{11}$$

The inverse PPI transform (i.e., converting ϕ_i to x_i) is

$$x_i \equiv F_i^{-1}\left(\, \Phi(\phi_i)\, \right). \tag{12}$$

130   The PPI transform generalizes the CAC2020 algorithm to handle non-Gaussian forecast ensembles. The resulting PESE-GC procedure has four stages and is illustrated in Fig. 3. These four stages are:

1. For each model state variable, fit a user-specified univariate distribution to that model variable in the forecast ensemble (i.e., marginal distribution fitting).

2. For each model state variable, apply the PPI transform [Eq. (11)] using that variable's fitted distribution to convert the
135   forecast ensemble of that model variable into forecast probits.

3. For each model state variable, adjust the mean and variance of that variable's forecast ensemble probits to zero and unity, respectively (explained in section 2.3), and then apply the CAC2020 algorithm on that variable's forecast probits to generate virtual probits.
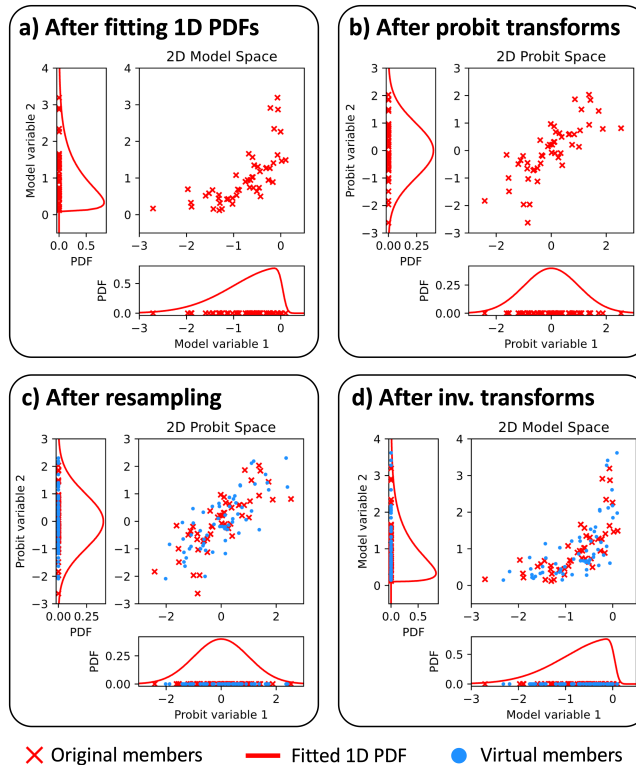
Figure 3. Illustrations of PESE-GC's 4-stage algorithm. Panels a, b, c, and d respectively show the aftermath of stages 1, 2, 3, and 4. The details of these stages are described in Section 2.2. Note that the CAC2020 algorithm is applied in stage 3 (i.e., between panels c and d).

    4. For each model state variable, apply the inverse PPI transform [Eq. (12)] with stage 1's fitted distribution on that variable's virtual probits to generate that variable's virtual ensemble.

To be clear, the CAC2020 algorithm is implemented and used in stage 3.

    Note that this four-stage assumes that the multivariate forecast distribution is Gaussian in probit space. This assumption arises from the use of a Gaussian resampling algorithm (the CAC2020 algorithm) to generate virtual probits. This assumption is equivalent to assuming that the multivariate forecast distribution has a Gaussian copula. As such, this four-stage procedure is called "Probit-space Ensemble Size Expansion for Gaussian Copulas".

    PESE-GC's four-stage procedure is attractive for geophysical EnsDA for several reasons. Aside from the fact that it can generate non-Gaussian virtual members, PESE-GC can be implemented in an embarrassingly parallel fashion (every loop over the model state variables is embarrassingly parallel). Furthermore, PESE-GC is likely affordable for geophysical EnsDA because the CAC2020 algorithm (stage 3) is efficient (see section 2.1).

    Note that the quality of the virtual members depends on the distributions the user selects in step 1 of PESE-GC. This will be discussed later in Section 3.

2.3 On the mean and variance adjustments in PESE-GC's step 3

PESE-GC requires forecast ensemble probits with zero mean and unity variance. Otherwise, the resulting virtual members will disobey the marginal distributions fitted in PESE-GC's step 1. However, because the forecast ensemble size is finite, the forecast ensemble's probits may have non-zero mean and non-unity variance. To illustrate, suppose PESE-GC is applied on 5 univariate forecast ensemble members (red crosses in Fig. 2c) and a Gaussian-tailed rank histogram distribution (Anderson, 2010) is fitted to those 5 members in PESE-GC's step 1. Applying the PPI transform (PESE-GC's step 2) results in forecast probits with mean zero and a variance of approximately 0.561. 10^7 virtual probits are then generated from these forecast probits using the CAC2020 algorithm, and the inverse PPI transform is applied to generate the virtual members (PESE-GC's step 4). The histogram-estimated PDF of the virtual members (blue curve in Fig. 2c) disagrees with the fitted (i.e., desired) Gaussian-tailed rank histogram PDF.

This problematic disagreement is resolved by adjusting the forecast probits' mean and variance to zero and unity (respectively) before generating the virtual members. Suppose the probit of the n-th forecast member for model variable i is $\phi_{i,n}$ and the pre-adjusted prior ensemble probit's sample mean and sample variance are μ and σ^2 respectively. The adjustment process is simply

$$\phi_{i,n} \leftarrow \frac{\phi_{i,n} - \mu}{\sigma} \quad \forall n = 1, 2, \ldots, N_E. \tag{13}$$

The impact of this adjustment is illustrated in Fig. 2: the virtual members' histogram-estimated PDF (blue curve in Fig. 2d) now matches the Gaussian-tailed rank histogram PDF in PESE-GC's step 1.

3 Conceptual exploration of PESE-GC's EnsDA impacts

To understand the influence of PESE-GC on EnsDA, consider a joint model-observation space formulation of Bayes' rule:

$$p(\boldsymbol{\psi}|\boldsymbol{y^o}) \propto p(\boldsymbol{y^o}|\boldsymbol{\psi}) \, p(\boldsymbol{\psi}) \tag{14}$$

where $\boldsymbol{y^o}$ is an N_y-dimensional vector of observation values,

$$\boldsymbol{\psi} \equiv \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{h}(\boldsymbol{x}) \end{bmatrix}, \tag{15}$$

$\boldsymbol{h}(\boldsymbol{x})$ is the observation operator, $p(\boldsymbol{y^o}|\boldsymbol{\psi})$ is the observation likelihood function, and $p(\boldsymbol{\psi})$ is the prior probability density function (pdf). The goal of EnsDA is to obtain a posterior ensemble that represents the posterior distribution $p(\boldsymbol{\psi}|\boldsymbol{y^o})$. If PESE-GC alters $p(\boldsymbol{y^o}|\boldsymbol{\psi})$ and/or $p(\boldsymbol{\psi})$, then PESE-GC will influence EnsDA. As such, there are two potential mechanisms for PESE-GC to influence EnsDA: 1) alterations to the observation likelihood used by the EnsDA algorithm, and 2) alterations to the prior distribution used by the EnsDA.

This section explores the influence of PESE-GC on EnsDA through those two mechanisms. A bivariate example will be used to illustrate those mechanisms. Suppose a scalar forecast model variable x has a skewed-normal forecast distribution and
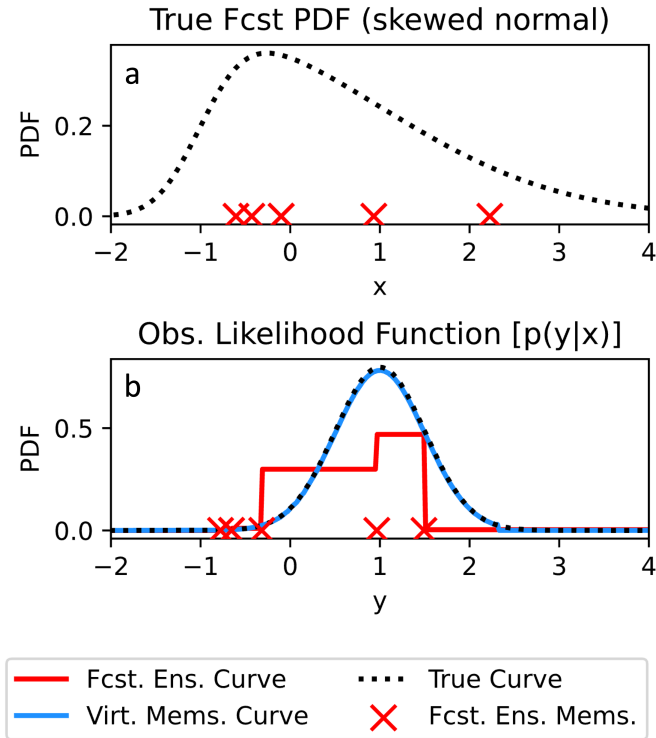
Figure 4. Plots of the true forecast PDF (a) and observation likelihood functions (b). The 5 forecast ensemble x values are indicated with red crosses in (a), and the 5 forecast ensemble y values are indicated with red crosses in (b). The red curve in (b) indicates the piecewise-approximated observation likelihood function used by a rank histogram filter that only uses the 5 forecast members. With PESE-GC, the rank histogram filter's piecewise-approximated observation likelihood function (blue curve) is improved.

5 samples are drawn from this distribution (illustrated in Figure 4a). A signed square-root function $h(x)$

$$h(x) \equiv \text{sign}(x) \sqrt{|x|} \tag{16}$$

will be used as the observation operator, and let y denote observation values. 10,000 virtual members will be generated by PESE-GC in this bivariate example. Note that the true forecast distribution of x is the previously mentioned skewed-normal distribution (illustrated in Figure 4a), and the true forecast distribution of $h(x)$ is estimated by applying $h(x)$ on 1,000,000 samples of x drawn from the true forecast distribution of x.

3.1 Mechanism 1: PESE-GC improves ensemble-based representation of observation likelihood function

For certain EnsDA algorithms that employ ensemble-based representations of the observation likelihood function, PESE-GC can improve those representations (impact mechanism 1). Two EnsDA algorithms will be considered: 1) the Rank Histogram Filter (RHF; Anderson (2010)), and 2) the serial stochastic Ensemble Kalman Filter (serial stochastic EnKF; Burgers et al. (1998)). In the case of the RHF, an ensemble-based piecewise approximation to the observation likelihood function is used
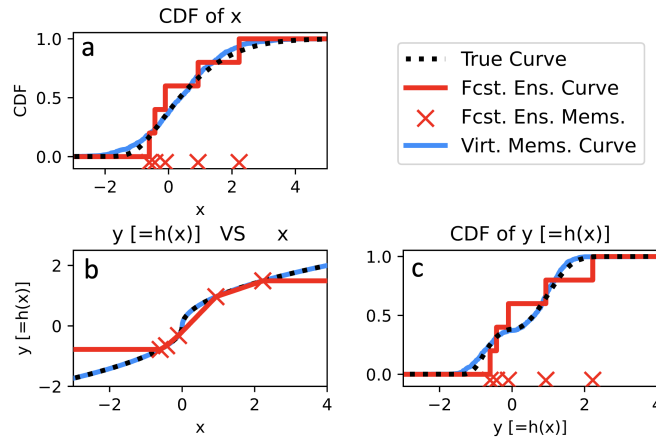
Figure 5. Bivariate example demonstrating the impacts of drawing virtual members from an informative fitted marginal (normal distribution). Panel a shows the empirical CDFs of x from the initial members and virtual members. The estimated relationships between x and y (obtained by passing the initial members and virtual members through $h(x)$) are displayed in panel b. Finally, panel c shows the empirical CDFs of the initial members and virtual members for variable y. The true CDFs and x-y relationships are plotted in panels b and c with dotted black lines.
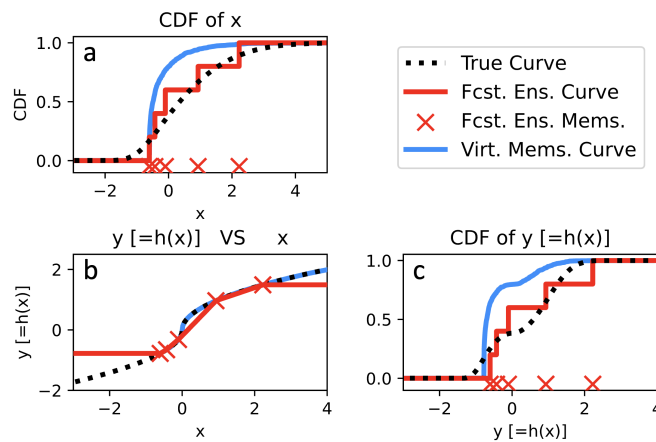
Figure 6. Bivariate example demonstrating the impacts of drawing virtual members from a misinformed fitted marginal (gamma distribution). The panels here are similar to Fig. 5.

(illustrated in Fig. 4b). The accuracy of that piecewise approximation depends on the ensemble size. When small forecast ensembles are used, that piecewise approximation is crude (red curve in Figure 4b). Since PESE-GC increases ensemble size, PESE-GC refines that piecewise representation (blue curve in Figure 4b). In the absence of other factors, this refinement will

195   improve the performance of the RHF.

Impact mechanism 1 also manifests for the serial stochastic EnKF. To see that, consider a situation with two observations, and recall that the serial stochastic EnKF uses random draws from a univariate Gaussian distribution to represent the likelihood function (one draw per ensemble member). For small ensembles, only a few of those random draws are made. In other words, there are sampling errors in representing the likelihood function. The ensemble statistics resulting from assimilating the first observation are thus degraded by those sampling errors. This degradation then affects the assimilation of the second observation. The assimilation of more than two observations compounds such sampling issues. Since PESE-GC increases the ensemble size, more draws from the likelihood function are made, thus suppressing sampling errors. As such, in the absence of other factors, PESE-GC will improve the performance of the serial stochastic EnKF.

Note that many EnsDA algorithms are immune to impact mechanism 1. The deterministic variants of the EnKF (Bishop et al., 2001; Whitaker and Hamill, 2002; Anderson, 2003; Tippett et al., 2003; Sakov et al., 2008) and particle filters (Gordon et al., 1993; Snyder et al., 2008; Poterjoy, 2016; van Leeuwen et al., 2019) are immune to impact mechanism 1 because no ensemble-based representation of the likelihood function is used. Also, the stochastic EnKF that assimilates all observations simultaneously (all-at-once stochastic EnKF) is immune to this effect – the chain of events described in the previous paragraph will not occur for the all-at-once stochastic EnKF. Finally, EnsDA systems that run ensembles of variational data assimilation methods (e.g., the ECMWF) are also immune to impact mechanism 1. This immunity is due to the same reason as the all-at-once stochastic EnKF.

3.2 Mechanism 2: PESE-GC influences ensemble statistics

If the user specifies appropriate marginals in PESE-GC's stage 1, then PESE-GC will improve the ensemble statistics used by EnsDA algorithms (mechanism 2). This will be illustrated using the bivariate 5-member example discussed near the start of Section 3. Suppose the user knows that the prior marginal distribution of x is close to Gaussian. Thus, a Gaussian distribution is fitted to the forecast x values in PESE-GC's stage 1 (Section 2.2; Fig. 5a). Applying PESE-GC generates 10,000 virtual members. Figure 5 indicates that the virtual members have better ensemble statistics than the forecast ensemble. This improvement in ensemble statistics is a combination of 1) the user selecting an appropriate marginal for x, and 2) the nonlinear $h(x)$ is evaluated more often (i.e., better sampled) with a large ensemble. As such, if the user specifies appropriate marginals in PESE-GC's stage 1, PESE-GC will likely improve the performance of EnsDA algorithms.

An important caveat is that if the user selects misinformed marginal distributions, then PESE-GC may degrade the ensemble statistics used by EnsDA algorithms. To illustrate, suppose the user fits a shifted gamma distribution (Cheng and Amin, 1983) to the 5 forecast x values. This distribution has three parameters: shape, scale and location. Since only 5 forecast values are used to fit three parameters, the fitted parameters' sampling errors are severe. Applying PESE-GC with this badly estimated shifted gamma distribution results in virtual ensemble statistics that are worse than those of 5 forecast x values (Fig. 6). In such situations, the performance of EnsDA will likely be degraded by PESE-GC. As such, the selection of marginals to use with PESE-GC must be done with care.

In the absence of knowledge about the model variables' prior marginal distribution, users can use non-parametric marginal distributions with PESE-GC. Such distributions include the Gaussian-tailed rank histogram distribution (Anderson, 2010) and

kernel distibutions (Anderson and Anderson, 1999). Choosing such distributions may not improve the model-space ensemble statistics. However, because PESE-GC increases the number of observation operator evaluations, the observation-space ensemble statistics and the ensemble covariance/copula between observation and model quantities can be improved.

Note that when linear observations are assimilated, PESE-GC with Gaussian marginals will not change the performance of deterministic variants of the EnKF (Bishop et al., 2001; Whitaker and Hamill, 2002; Anderson, 2003; Tippett et al., 2003; Sakov et al., 2008). This is because the expanded ensemble will have the same joint-space mean and covariance as the forecast ensemble.

4   Tests with Lorenz 1996 model

4.1   Setup of experiments

This section explores the impacts of PESE-GC on the performance of EnsDA using perfect model Observing System Simulation Experiments (OSSEs) with the Lorenz 1996 model (L96 model; Lorenz (2006)). The Data Assimilation Research Testbed (DART; https://github.com/NCAR/DART; Anderson et al. (2009)) is used in this exploration and PESE-GC has been implemented into DART.

The L96 model uses 40 variables (i.e., 40 grid points in a ring), a forcing parameter value of 8 (i.e., $F = 8$), and a time-step of 0.05 L96 time units. The L96 time unit is henceforth referred to as τ. Forward time integration of the model is done via the Runge-Kutta fourth-order integration scheme. Every OSSE experiment runs for 5,500 cycles. Initial nature run states and the initial ensemble members are drawn from the L96's climatology.

Note that the 0.05τ time step has been arbitrarily taken to be either 1 hour (Anderson, 2019, 2023) or 6 hours (Lorenz, 2006; Kurosawa and Poterjoy, 2021). To avoid potential confusion, this paper will avoid converting L96 time units to hours. In other words, all results will be displayed and discussed in terms of L96 time units.

In all experiments, there are 40 observations. Their observation locations are fixed throughout this study. Supposing that the model grid points have locations $0.025m$, $m = 1, 2, \ldots, 40$, each site location is a random draw from a uniform distribution between 0 and 1.

PESE-GC's impacts are examined using EnsDA experiments that are conducted with four N_Es, five cycling intervals, three post-PESE-GC ensemble sizes, three observation types, four EnsDA algorithms, and with and without PESE-GC, a total of $4 \times 5 \times 3 \times 3 \times 4 \times 2 = 1440$ configurations. The four N_Es are 10, 20, 40, and 80, and the five cycling intervals are 0.05τ, 0.10τ, 0.15τ, 0.30τ, and 0.60τ.

The PESE-GC-expanded ensemble sizes are specified in terms of factors: 5, 10 and 20 times the forecast ensemble size. For example, if $N_E = 10$, a PESE-GC factor of 10 means that the expanded ensemble has 100 members (i.e., $N_V = 90$ virtual members are created).

260   Supposing the k-th observation site has location ℓ_k, the observation operators for the three observation types are:

$$h_{IDEN}\left(\boldsymbol{x};\ell_k\right) \equiv x_{\ell_k}, \tag{17}$$

$$h_{SQRT}\left(\boldsymbol{x};\ell_k\right) \equiv sgn\left(x_{\ell_k}\right)\sqrt{|x_{\ell_k}|}, \tag{18}$$

$$h_{SQUARE}\left(\boldsymbol{x};\ell_k\right) \equiv sgn\left(x_{\ell_k}\right)\left(x_{\ell_k}\right)^2 \tag{19}$$

where $\boldsymbol{x}$ is the L96 model's 40-element state vector, x_{ℓ_k} is the model variable linearly interpolated to location ℓ_k, and $sgn\left(x_{\ell_k}\right)$

265   returns the sign of x_{ℓ_k}. Observations created using $h_{IDEN}\left(\boldsymbol{x};\ell_k\right)$, $h_{SQRT}\left(\boldsymbol{x};\ell_k\right)$ and $h_{SQUARE}\left(\boldsymbol{x};\ell_k\right)$ will be respectively referred to as IDEN observations, SQRT observations and SQUARE observations. Every observation is created by applying its corresponding observation operator on a nature run state, and then perturbing the output with a random draw from $\mathcal{N}\left(0,\sigma^2\right)$. Following Anderson (2020), the chosen σ^2 for the IDEN, SQRT and SQUARE observations are 1.0, 0.25, and 16, respectively.

     The four EnsDA algorithms tested with PESE-GC are:

270   1. the Ensemble Adjustment Kalman Filter (EAKF; Anderson (2003))

     2. the serial stochastic EnKF with sorted observation increments (EnKF; Burgers et al. (1998))

     3. the Rank Histogram Filter with linear regression (RHF; Anderson (2010))

     4. the Rank Histogram Filter with probit regression (PR; Anderson (2023))

To be clear, the RHF algorithm first employs the rank histogram filter to generate observation increments and then uses linear

275   regression to convert observation increments into model increments. The PR algorithm is similar to the RHF algorithm, except that probit regression is used to convert observation increments into model increments.

     For each EnsDA algorithm, only 1 set of marginals are used with PESE-GC. When PESE-GC is used with the EAKF, EnKF or RHF algorithm, Gaussian marginals are selected for all 40 model variables. PESE-GC with Gaussian marginals is identical to the CAC2020 algorithm. In other words, for the EAKF, EnKF and RHF experiments, the virtual ensemble members follow

280   multivariate Gaussian distributions. For the PR algorithm, the Gaussian-tailed rank histogram is selected as the marginal for every one of the 40 model variables. This means the PR experiments' virtual ensemble members follow multivariate non-Gaussian distributions. Future work can investigate the impacts of using PESE-GC with Gaussian-tailed rank histograms (or kernel density estimates) with the EAKF, EnKF and RHF.

     Each of the 1440 configurations is trialed 36 times. These trials are enumerated (Trial 1, Trial 2, and so forth). All experiments

285   with the same trial number and N_E share the same nature run and initial forecast ensemble states. For example, the following two EnsDA experiments have the same initial nature run and initial forecast ensemble: 1) Trial 10 using IDEN observations, 20 forecast ensemble members, 0.15τ cycling period, EAKF, and PESE-GC with an expansion factor of 20, and 2) Trial 10 using SQRT observations, 20 forecast ensemble members, 0.60τ cycling period, RHF, and PESE-GC with an expansion factor of 5. Note that experiments with different trial numbers have different nature runs and initial forecast ensemble states.

290   The Gaspari-Cohn fifth-order rational function (Gaspari and Cohn, 1999) is used to localize EnsDA increments. For each combination of trial and configuration, 17 localization half-radii are tested: $0.075 \times 1.3^0, 0.075 \times 1.3^1, 0.075 \times 1.3^2, 0.075 \times$

$1.3^3, \ldots, 0.075 \times 1.3^{14}, 0.075 \times 1.3^{15}$, and infinity. To select the optimal localization half-radius for a given combination of trial and configuration, the root-mean-square error (RMSE) of the forecast ensemble mean is used. The RMSE for a particular cycle is

$$
\text{RMSE} \equiv \sqrt{\frac{1}{40} \sum_{i=1}^{40} \left(\overline{x_i^f} - x_i^t \right)^2} \tag{20}
$$

where $\overline{x_i^f}$ and x_i^t are, respectively, the forecast ensemble mean state and nature run state at model grid point i. The RMSE of the first 500 out of 5,500 EnsDA cycles are discarded. The localization half-radius that results in the smallest cycle-averaged RMSE (i.e., averaged from cycles 501 to 5,500) is selected as the optimal localization half-radius. Note that for the same configuration, the optimal localization half-radius can vary with the trial number.

The inflation scheme used here is identical to the one used by Anderson (2023). The adaptive prior inflation algorithm of Anderson (2009) is used with an inflation lower bound of 1 (no deflation), an upper bound of 2, a fixed inflation standard deviation of 0.6, and an inflation damping of 0.9. While manually tuning a homogeneous inflation factor or a relaxation-to-posterior-spread (RTPS; Whitaker and Hamill (2012)) relaxation factor may give smaller RMSEs, an adaptive inflation approach is chosen to reduce the computational cost of this study. This study already runs 881,280 ($1,440 \times 36 \times 17$) combinations of configurations (1,440), trials (36), and localization half-radii (17), and each combination is run for 5,500 cycles.

4.2 Metric to assess PESE-GC's impact on EnsDA

The impacts of PESE-GC on EnsDA are assessed using the relative difference between cycle-averaged RMSEs (Eq. (20)) with PESE-GC and cycle-averaged RMSEs without PESE-GC. The cycle averaging is done from cycles 501 to 5,500. To define this RMSE relative difference, suppose an arbitrary configuration of N_E, cycling interval, PESE-GC factor, observation type, and EnsDA algorithm is denoted by $\boldsymbol{\xi}$. Let $\overline{\text{RMSE}(\boldsymbol{\xi}, r, \text{True})}$ denote the cycle-averaged RMSE of the r-th trial of configuration $\boldsymbol{\xi}$ with PESE-GC used. Furthermore, let $\overline{\text{RMSE}(\boldsymbol{\xi}, r, \text{False})}$ denote the cycle-averaged RMSE of the r-th trial of configuration $\boldsymbol{\xi}$ without using PESE. The relative difference between $\overline{\text{RMSE}(\boldsymbol{\xi}, r, \text{True})}$ and $\overline{\text{RMSE}(\boldsymbol{\xi}, r, \text{False})}$ is defined as

$$
\Delta\text{RMSE}(\boldsymbol{\xi}, r) \equiv \frac{\text{RMSE}(\boldsymbol{\xi}, r, \text{True}) - \text{RMSE}(\boldsymbol{\xi}, r, \text{False})}{\left\langle \overline{\text{RMSE}(\boldsymbol{\xi}, \text{False})} \right\rangle}. \tag{21}
$$

Here, the denominator is the trial-averaged (indicated by angled brackets) cycle-averaged RMSE of EnsDA run with the same $\boldsymbol{\xi}$ as in the numerator, but with PESE-GC unused. For readability, $(\boldsymbol{\xi}, r)$ is omitted from the rest of this paper. Most importantly, a negative ΔRMSE indicates that PESE-GC improves the performance of EnsDA, and a positive value of ΔRMSE indicates that PESE-GC degrades the performance of EnsDA.

Only statistically significant trial-averaged ΔRMSE (henceforth, $\langle\Delta\text{RMSE}\rangle$) will be discussed in this paper. A $\langle\Delta\text{RMSE}\rangle$ is considered statistically significant if its two-tailed z-test p value is smaller than 1%. These statistically significant $\langle\Delta\text{RMSE}\rangle$ values are plotted in Figs. 7, 8, and 9.

Impacts of PESE on RMSE (IDEN obs)
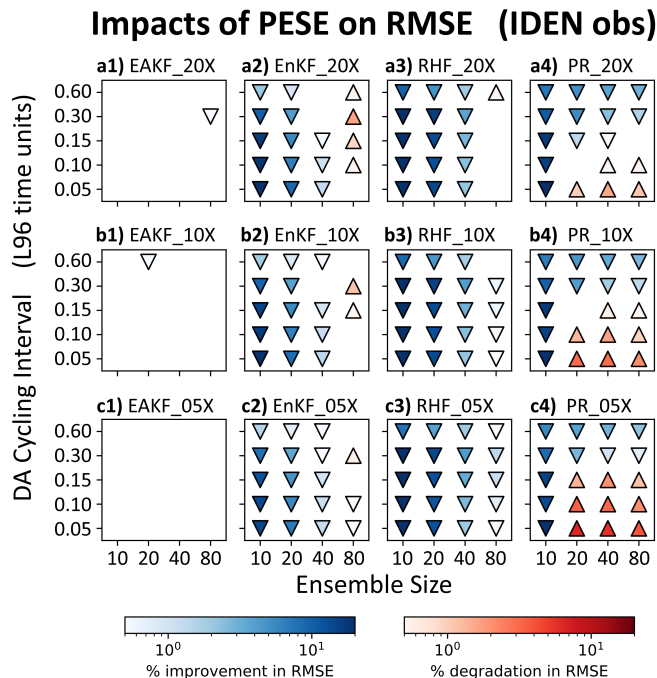
Figure 7. Statistically significant $\langle\Delta\text{RMSE}\rangle$ values (two-tailed $p < 0.01$) for pairs of PESE-GC-using and PESE-GC-omitting experiments that assimilate IDEN observations. The PESE-GC-using EnsDA experiments in panels a1, a2, a3 & a4 expanded their forecast ensembles 20-fold, the PESE-GC-using EnsDA experiments in panels b1, b2, b3 & b4 expanded their forecast ensembles 10-fold, and the PESE-GC-using EnsDA experiments in panels c1, c2, c3 & c4 expanded their forecast ensembles 5-fold. Panels a1, b1 and c1 show $\langle\Delta\text{RMSE}\rangle$ for EnsDA experiments using the EAKF, panels a2, b2 and c2 show $\langle\Delta\text{RMSE}\rangle$ for EnsDA experiments using the EnKF, panels a3, b3 and c3 show $\langle\Delta\text{RMSE}\rangle$ for EnsDA experiments using the RHF, and panels a4, b4 and c4 show $\langle\Delta\text{RMSE}\rangle$ for EnsDA experiments using the PR. The acronyms are defined in Section 4.1. Downward triangles indicate $\langle\Delta\text{RMSE}\rangle < 0$ and upward triangles indicate $\langle\Delta\text{RMSE}\rangle > 0$.

Before proceeding, note that PESE-GC with Gaussian marginals only negligibly changes the performance of the EAKF with IDEN observations (Fig. 7a1, b1 & c1). This negligibility is predicted in Section 3.2. Any impact of PESE-GC on the performance of these experiments is likely due to rounding errors associated with the use of finite precision arithmetic.

4.3 Impacts of using 20-fold PESE-GC on EnsDA

325 This study first examines the $\langle\Delta\text{RMSE}\rangle$ for a PESE-GC factor of 20 (panels a1, a2, a3 and a4 in Figs. 7, 8 and 9). The focus is on identifying common patterns in $\langle\Delta\text{RMSE}\rangle$ and explaining them through the lens of the two mechanisms laid out in Section 3. The variation of $\langle\Delta\text{RMSE}\rangle$ with different PESE-GC factors (remaining panels in Figs. 7, 8, and 9) will be discussed in Section 4.4.

The first common $\langle\Delta\text{RMSE}\rangle$ pattern in the 20-fold PESE-GC situations is that PESE-GC generally improves EnsDA per-
330 formance (i.e., $\langle\Delta\text{RMSE}\rangle < 0$) when N_E is 10 or 20 (panels a1, a2, a3 and a4 in Figs. 7, 8 and 9). This is because either one
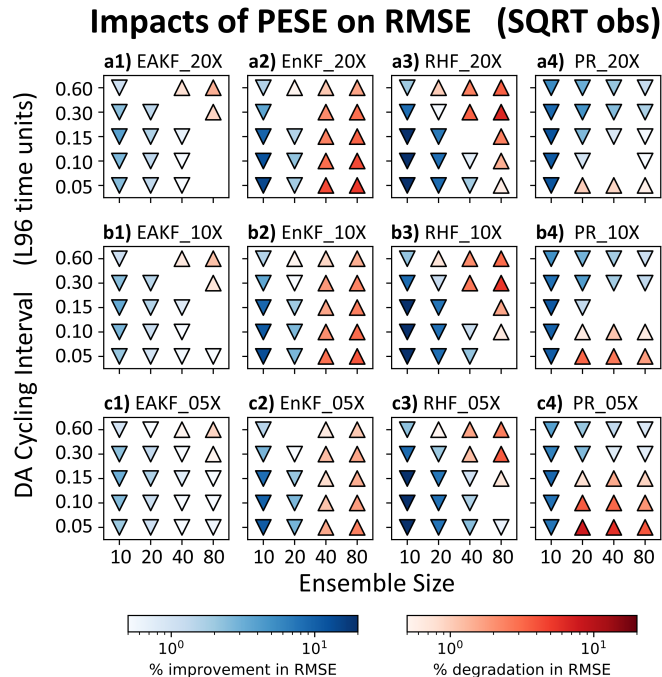
Figure 8. Similar to Fig. 7, except that SQRT observations are assimilated.

or both mechanisms described in Section 3 are acting to improve RMSEs. First, for the PR, RHF and EnKF experiments, the performance improvements are partly (if not entirely) because PESE-GC improves the ensemble-based representation of the likelihood function (i.e., mechanism 1; Section 3.1). For the PR experiments with IDEN observations, mechanism 1 is likely the sole reason for the improved performance. Second, the PESE-GC-induced RMSE reductions in all 10/20 forecast member

335 experiments with either SQRT or SQUARE observations are partly due to improved sampling of the observation operator (i.e., mechanism 2 described in Section 3.2).

The RMSE reductions seen in the 10/20-member EAKF, EnKF and RHF experiments are also partly due to improved ensemble statistics (i.e., mechanism 2; Section 3.2). This is because the L96 model's forecast statistics tend to be close to Gaussian (e.g., Chan et al. (2020)). Applying PESE-GC with Gaussian marginals thus improves the model variables' prior

340 ensemble statistics, therefore improving the performance of the EAKF, EnKF and RHF experiments.

The second common pattern in the 20-fold PESE-GC experiments is that with increasing N_E, PESE-GC's RMSE impacts can go from improving ($\langle\Delta\text{RMSE}\rangle < 0$) to degrading ($\langle\Delta\text{RMSE}\rangle > 0$). This pattern is likely related to both mechanisms in Section 3. First, for the EnKF, RHF and PR, the ensemble-based representation of the observation likelihood function improves with increasing N_E. This improvement implies there is less room for PESE-GC to improve that representation. As

345 such, mechanism 1 weakens with increasing N_E, thus reducing PESE-GC-induced EnsDA performance gains for the EnKF, RHF and PR EnsDA algorithms.
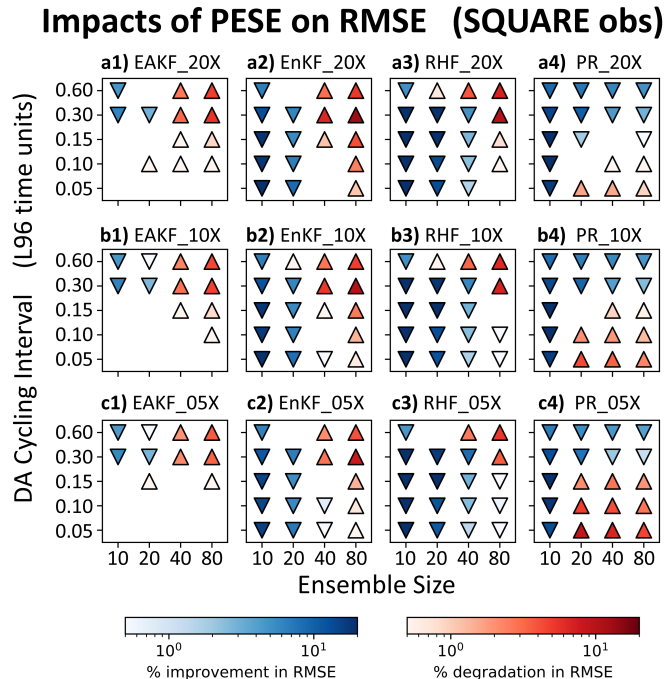
Figure 9. Similar to Fig. 7, except that SQUARE observations are assimilated.

For the EAKF, EnKF and RHF experiments, impact mechanism 2 also contributes to the worsening of PESE-GC's RMSE impacts with increasing N_E. In the act of choosing Gaussian marginal distributions for PESE-GC, the user implicitly assumes that the true forecast PDF is Gaussian. With increasing N_E, imperfections in this Gaussian assumption become increasingly evident. The impacts of PESE-GC on the observation space prior statistics can thus go from improving to degrading with increasing N_E. This change likely contributes to the worsening of PESE-GC's RMSE impacts with increasing N_E.

A third common pattern is that with longer cycles, the PESE-GC's RMSE impacts on the EAKF, EnKF and RHF degrades (i.e., $\langle \Delta \text{RMSE} \rangle$ goes from either negative to zero, zero to positive, or negative to positive). This pattern is likely due to increasing non-Gaussianity in the forecast ensemble's statistics with longer cycles. The choice to use Gaussian marginals in these experiments' PESE-GC thus becomes increasingly inappropriate, meaning that impact mechanism 2 increasingly degrades the performance of EnsDA.

The fourth common pattern is that in the PR experiments, for $N_E \geq 20$, the 20-fold PESE-GC's impact improves with longer cycling interval (Figs. 7a4, 8a4 and 9a4). A plausible explanation relates to PR's usage of a piecewise approximation to the observation likelihood function (henceforth, the "piecewise approximation"). This approximation is more accurate when more ensemble members sample the regions of the observation likelihood function where that function varies strongly. However, with increasing forecast ensemble variance, those regions tend to be less sampled by forecast ensemble members, thus degrading the piecewise approximation. Since longer cycling intervals increase forecast ensemble variance, longer cycling intervals thus increase the room for PESE-GC to improve the piecewise approximation. Future work can test this explanation.

Note that the chain of events discussed in the previous paragraph likely occurs for the RHF experiments as well. Since the
RHF experiments do not exhibit the fourth common pattern, it is likely that the inappropriateness of the Gaussian marginals
used with PESE-GC overwhelms improvements introduced by refining the piecewise approximation.

4.4 Variations in PESE-GC's impacts with different amounts of ensemble expansion

This study now examines common patterns in how PESE-GC's impacts vary with PESE-GC expansion factors. The first
common pattern is that PESE-GC's impacts on the PR experiments tends to weaken with smaller PESE-GC factors (panels a4,
b4 and c4 in Figs. 7, 8 and 9). This pattern is likely caused by increasing sampling errors in the virtual members' statistics with
fewer virtual members (i.e., smaller PESE-GC factors).

The second common pattern is that, for $N_E \geq 40$, smaller PESE-GC factors tend to result in milder RMSE degradations
introduced by PESE-GC in the EAKF, EnKF and RHF experiments. Since these RMSE degradations are likely due to the
misinformed selection of Gaussian marginals with PESE-GC (see section 4.3), reducing the number of virtual members reduces
the amount of misinformation introduced by PESE-GC into the forecast statistics. These less misinformed forecast statistics
explain the second common pattern.

An interesting third common pattern is also visible in the EAKF, EnKF and RHF experiments: there are instances where
reducing PESE-GC factors 1) turns insignificant RMSE impacts into RMSE improvements (e.g., the lower left corner of
Figs. 8a1 and c1), or 2) turns RMSE degradations into RMSE improvements (e.g., upper right corner of Figs. 7a3 and c3).
To explain this third pattern, notice that these instances are associated with short cycling intervals (0.05–0.10τ) and shorter
cycling intervals are associated with increasingly Gaussian forecast PDFs. Based on these associations, a plausible explanation
is that the true forecast statistics only mildly deviate from Gaussian, but the forecast ensemble statistics are often far from
Gaussian. Introducing some Gaussian virtual members thus improves the ensemble statistics. However, if too many Gaussian
virtual members are introduced, the expanded ensemble statistics become too close to Gaussian. This "Goldilocks" explanation
can be tested in future work.

Most importantly, even with a mere 5-fold PESE-GC, PESE-GC improves the performance of EnsDA in three types of
situations. First, all EnsDA experiments involving small forecast ensemble sizes (10 members) are improved by PESE-GC.
Second, situations where using Gaussian marginals with PESE-GC improves ensemble statistics are also improved by PESE-
GC. This second type of situation occurs for the EAKF, EnKF and RHF experiments that either have 1) 20-40 ensemble
members and/or 2) cycling intervals that are 0.30τ or less. Third, PESE-GC improves the PR experiments for cycling intervals
that are 0.30τ or 0.60τ. This improvement is plausibly because with longer cycling intervals, PESE-GC better improves the
piecewise observation likelihood approximation used by the PR EnsDA algorithm (explained in Section 4.3). These PESE-
GC-introduced improvements are particularly encouraging because a geophysical EnsDA system is more likely able to afford
using 5-fold PESE-GC over 20-fold PESE-GC.
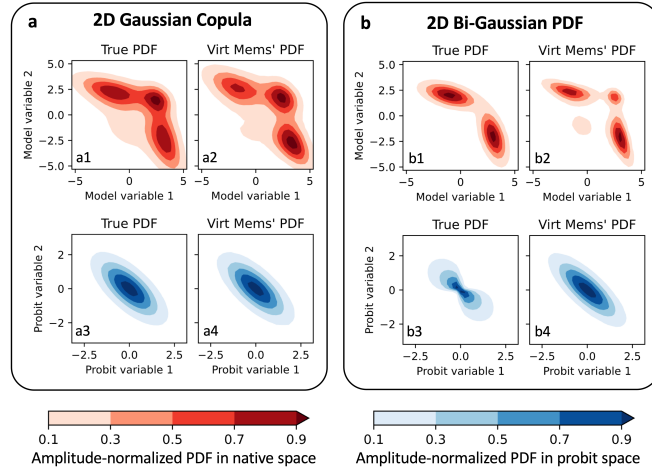
Figure 10. Two bivariate demonstrations of PESE-GC. In each demonstration, 100 initial members are drawn from a true bivariate PDF (a1, b1), Bi-Gaussian PDFs are fitted to each variable, and PESE-GC creates 1,000,000 virtual members. Panels a1 and b1 show the PDFs that the initial members are drawn from (i.e., the true bivariate PDFs), panels a2 and b2 show bivariate empirical PDFs estimated by the virtual members, panels a3 and b3 show the true bivariate PDF in probit space, and panels a4 and b4 show the virtual members' bivariate empirical PDFs in probit space. The two true bivariate are: a trimodal PDF with an underlying Gaussian copula (a1) and a bi-Gaussian PDF (panel d1). Note that the bi-Gaussian PDF's copula is not a Gaussian copula (b3). The two true PDFs are described in Section 5.1.

5   Discussion

5.1   PESE-GC assumes Gaussian copulas

The results presented in the previous section are encouraging. However, a caveat about PESE-GC needs discussion: PESE-GC assumes that the forecast distribution is a multivariate Gaussian distribution in probit space (i.e., Gaussian copula). If that assumption is violated (henceforth, "Gaussian copula assumption"), the virtual members will possess statistical artifacts.

Fig. 10a illustrates PESE-GC's ability to generate non-Gaussian virtual members for a situation where the Gaussian copula assumption holds. The true forecast multivariate PDF (Fig. 10a1) is created by applying two inverse PPI transforms on a bivariate Gaussian PDF. The two-dimension mean vector $\boldsymbol{\mu}$ and the 2×2 covariance matrix $\boldsymbol{\Sigma}$ of the bivariate Gaussian PDF are

$$\boldsymbol{\mu} \equiv \begin{bmatrix} 0 & 0 \end{bmatrix}^{\top}, \ \boldsymbol{\Sigma} \equiv \begin{bmatrix} 1 & 0.7 \\ 0.7 & 1 \end{bmatrix}. \tag{22}$$

The first inverse PPI transform is applied on the first variable (x_1) and the PDF it uses $(p(x_1))$ is

$$p(x_1) \equiv \frac{1}{2} G(x_1; -1, 2) + \frac{1}{2} G(x_1; 3, 1) \tag{23}$$

where $G(x_1; -1, 2)$ represents the Gaussian PDF with scalar mean -1 and standard deviation 2, and likewise for $G(x_1; 3, 1)$. The second inverse PPI transform is applied on the second variable (x_2) and the PDF it uses ($p(x_2)$) is

$$p(x_2) \equiv \frac{1}{2}G(x_2; 2, 1) + \frac{1}{2}G(x_2; -2, 2) \tag{24}$$

410   where $G(x_2; 2, 1)$ and $G(x_2; -2, 2)$ are defined similar to $G(x_1; -1, 2)$.

Since the forecast PDF in Fig. 10a1 has, by construction, a Gaussian copula (Fig. 10a3), PESE-GC can produce virtual members that approximately follow the forecast PDF. To demonstrate, 100 forecast members were drawn from the forecast PDF, and 1,000,000 virtual members were constructed using PESE-GC. The two marginal PDFs that are used in steps 1, 2 and 4 of PESE-GC are univariate Bi-Gaussian PDFs (fitted via maximum likelihood estimation in step 1). The histogram-estimated

415   PDFs of the virtual members (Fig. 10a2) and virtual probits (Fig. 10a4) are similar to the true forecast PDF (Fig. 10a1) and the true forecast's probit-space PDF (Fig. 10a3).

An example where the Gaussian copula assumption is violated is shown in Fig. 10b. Here, the forecast PDF (Fig. 10b1) is the following bivariate bi-Gaussian PDF

$$p(\boldsymbol{x}) = \frac{1}{2}G(\boldsymbol{x}; \boldsymbol{\mu_1}, \boldsymbol{\Sigma_1}) + \frac{1}{2}G(\boldsymbol{x}; \boldsymbol{\mu_2}, \boldsymbol{\Sigma_2}) \tag{25}$$

420   where

$$\boldsymbol{x} \equiv \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \ \boldsymbol{\mu_1} \equiv \begin{bmatrix} -1 \\ 2 \end{bmatrix}, \ \boldsymbol{\mu_2} \equiv \begin{bmatrix} 3 \\ -2 \end{bmatrix}, \tag{26}$$

$$\boldsymbol{\Sigma_1} \equiv \begin{bmatrix} 2 & -0.5 \\ -0.5 & 0.5 \end{bmatrix}, \ \boldsymbol{\Sigma_2} \equiv \begin{bmatrix} 0.5 & -0.5 \\ -0.5 & 2 \end{bmatrix}. \tag{27}$$

Applying PPI transforms on this bivariate bi-Gaussian forecast PDF reveals that the bi-Gaussian PDF violates the Gaussian copula assumption (Fig. 10b3). When PESE-GC is applied to generate 1,000,000 virtual members from 100 forecast members,

425   the virtual members' probit space bivariate PDF (Fig. 10b4) differs from the forecast PDF in probit space (Fig. 10b3). As such, the virtual members' bivariate PDF (Fig. 10b2) deviates from the true bivariate bi-Gaussian forecast PDF (Fig. 10b1; the virtual members have two small spurious modes).

Note that though the virtual members' PDF deviates from the forecast PDF, a strong similarity exists between the two PDFs. The two dominant modes of the virtual members' PDF are very similar to the bi-Gaussian forecast PDF. More generally, milder

430   violations of the Gaussian copula assumption will likely lead to milder spurious statistical features in the virtual members.

More importantly, PESE-GC's Gaussian copula assumption may not be problematic for geophysical EnsDA. Due to the high dimensionality of geophysical models and small forecast ensemble sizes, it is difficult to identify the family of the multivariate forecast distributions in probit space. In other words, the forecast ensemble's statistics in probit space are likely indistinguishable from a multivariate Gaussian. This indistinguishability permits assuming Gaussian copulas. Future work can investigate

435   this possibility.

5.2 Impacts of PESE-GC on the computational cost of EnsDA

It is also important to discuss the impacts of PESE-GC on the EnsDA process (i.e., the forecast step and analysis step). Since the virtual members are deleted before running forecast models (Figure 1), PESE-GC does not change the forecast step's computational cost. However, PESE-GC will increase the computational cost of the analysis step because 1) the number of observation operator calls scales linearly with the ensemble size, and 2) the EnsDA filter algorithm's (e.g., the EAKF algorithm) computational complexity scales with the ensemble size. Supposing the computational cost of the EnsDA filter algorithm scales linearly with ensemble size, then the computational cost of the EnsDA analysis step scales linearly with the number of ensemble members created by PESE-GC.

However, the increase in the computational cost associated with PESE-GC is likely far more affordable than running a larger forecast ensemble. This is because the computational cost of the forecast step often accounts for $> 80\%$ of the overall computational cost of the EnsDA process and the analysis step accounts for the remaining $< 20\%$. Consider the situation where PESE-GC quintuples the ensemble size. The overall computational cost of the EnsDA process will only increase by $< 80\%$. In contrast, if the forecast ensemble size is quintupled, then the overall computational cost of the EnsDA process increases by 400%. As such, PESE-GC is an alluring alternative to increasing the forecast ensemble size.

6 Summary and future work

In this study, an efficient and embarrassingly parallel algorithm to increase ensemble sizes, PESE-GC, is formulated. PESE-GC generalizes the efficient and embarrassingly parallel Gaussian resampling algorithm of Chan et al. (2020) to handle non-Gaussian forecast distributions. This handling of non-Gaussian forecast distributions means PESE-GC is highly flexible. Furthermore, PESE-GC provides an avenue for users to use their knowledge of the forecast statistics to improve EnsDA – users can choose to draw virtual members using marginal distribution families (e.g., Gaussian and gamma distribution families) that they think are good approximations to the true forecast marginal distributions. If that knowledge is unavailable, users can choose to use non-parameteric marginal distributions (e.g., Gaussian-tailed rank histogram distributions).

Three mechanisms are then identified for PESE-GC to influence the performance of EnsDA. First, for EnsDA methods like the serial stochastic EnKF and the rank histogram filter, PESE-GC improves the representation of the observation likelihood function. Second, by expanding the number of ensemble members, PESE-GC increases the sampling of the observation operator. This increased sampling improves the forecast observations' PDF. Finally, when users use PESE-GC with informative marginal distribution families, the forecast observations' statistics are improved.

The impacts of PESE-GC on the performance of EnsDA are explored using the L96 model, a variety of observation systems and a variety of EnsDA algorithms. Results indicate that PESE-GC generally improves the performance of EnsDA when 1) the forecast ensemble size is small (10 members), 2) the marginal distribution families used with PESE-GC are informative, and/or 3) PESE-GC improves the representation of the observation likelihood function (the PR experiments in sections 4.3 and 4.4). It is particularly encouraging that many of these improvements are retained even with a modest amount of ensemble expansion (5-fold expansion).

There are two general areas for future work with PESE-GC. The first area is to move PESE-GC towards geophysical models (EnsDA or forecast postprocessing). To do so, PESE-GC needs to be first tested with ensemble members created by geophysical models (e.g., Weather Research and Forecasting model; Skamarock et al. (2008)). It will be particularly interesting to see if the virtual members have realistic meteorological structures (e.g., convective clouds with supporting circulations). Then, PESE-GC can be tested using geophysical EnsDA and/or postprocessing system. If PESE-GC does improve the performance of geophysical EnsDA/postprocessing, a comparison between PESE-GC and other ensemble expansion methods is warranted.

Another general area for future work is to develop the PESE-GC algorithm further. First, given the importance of localization in practical EnsDA, future work can and should explore inserting localization into PESE-GC. Second, the validity of PESE-GC's Gaussian copula assumption can be assessed in the context of geophysical modelling and forecasting. If the Gaussian copula assumption is inappropriate, then non-parametric methods to generate virtual probits can be explored. Third, methods to detect the usage of misinformed parametric marginal distribution families deserve exploration. One possible detection method is to employ hypothesis testing on the marginal distributions. For example, if Gaussian distributions are selected for PESE-GC, then the Shapiro-Wilk test can be applied on the forecast ensemble to determine if the selection is misinformed (e.g., Kurosawa and Poterjoy (2023)). Finally, the use of non-Gaussian marginals with PESE-GC may alter ensemble covariances between model variables. This possibility deserves future investigation.

The computational cost of running geophysical models will continue increasing in the coming years (higher spatial resolution, shorter time steps, more complex parameterization schemes, etc). Geophysical EnsDA groups will continue to grapple with the challenge of balancing the computational costs of increasing the number of forecast ensemble members and the computational costs of using more realistic geophysical models. If ensemble expansion methods can provide much of the benefits of a larger forecast ensemble size at a fraction of the cost, these methods will enable EnsDA groups to employ more realistic geophysical models.

Code availability. The codes used in this study is publicly available in Chan (2023). These codes include 1) Python scripts used to generate the conceptual illustrations, 2) an implementation of PESE-GC into DART, and 3) Python and Bash scripts used to run and evaluate this study's Lorenz 1996 experiments.

Author contributions. The author, Man-Yau Chan, created the PESE-GC algorithm, implemented PESE-GC into DART, designed and executed the experiments discussed in this publication, plotted every figure in this publication, and drafted and edited every version of this publication.

Competing interests. The author have no competing interests with this study.

Disclaimer. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Acknowledgements. The author is eternally grateful to Jeffrey L. Anderson and the National Center for Atmospheric Research's Data Assimilation Research Section for useful discussions and guidance. The author thanks Mohamad El Gharamti, and Craig Schwartz for helping to improve the explanation of the PESE-GC algorithm. Furthermore, the author thanks the participants of the International Symposium for Data Assimilation 2023 (ISDA), particularly Alberto Ortolani, for discussions that further clarified the author's thinking process and explanations. The author is also grateful to Yao Zhu and Christopher Hartman for checking the readability of this manuscript. Finally, the author would like to thank Olivier Talagrand (the Editor), Ian Groom (reviewer) and an anonymous reviewer for their thorough review of this manuscript and for their constructive feedback.

This study is supported by the Advanced Study Program Postdoctoral Fellowship at the National Center for Atmospheric Research (NCAR) and the Ohio State University. NCAR is sponsored by the National Science Foundation. All computations in this study are done on two NCAR computing clusters: Casper and Cheyenne. These clusters are managed by NCAR's Computational Information Systems Laboratory.

References

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

Anderson, J. L.: A Local Least Squares Framework for Ensemble Filtering, Monthly Weather Review, 131, 634–642, https://doi.org/10.1175/1520-0493(2003)131<0634:ALLSFF>2.0.CO;2, 2003.

Anderson, J. L.: Spatially and temporally varying adaptive covariance inflation for ensemble filters, Tellus, Series A: Dynamic Meteorology and Oceanography, 61 A, https://doi.org/10.1111/j.1600-0870.2008.00361.x, 2009.

Anderson, J. L.: A Non-Gaussian Ensemble Filter Update for Data Assimilation, Monthly Weather Review, 138, 4186–4198, https://doi.org/10.1175/2010MWR3253.1, 2010.

Anderson, J. L.: A nonlinear rank regression method for ensemble Kalman filter data assimilation, https://doi.org/10.1175/MWR-D-18-0448.1, 2019.

Anderson, J. L.: A marginal adjustment rank histogram filter for non-Gaussian ensemble data assimilation, Monthly Weather Review, 148, https://doi.org/10.1175/MWR-D-19-0307.1, 2020.

Anderson, J. L.: A Quantile-Conserving Ensemble Filter Framework. Part II: Regression of Observation Increments in a Probit and Probability Integral Transformed Space, Monthly Weather Review, 2023.

Anderson, J. L. and Anderson, S. L.: A Monte Carlo implementation of the nonlinear filtering problem to produce ensemble assimilations and forecasts, Monthly Weather Review, https://doi.org/10.1175/1520-0493(1999)127<2741:AMCIOT>2.0.CO;2, 1999.

Anderson, J. L., Hoar, T., Raeder, K., Liu, H., Collins, N., Torn, R., and Avellano, A.: The data assimilation research testbed a community facility, Bulletin of the American Meteorological Society, 90, 1283–1296, https://doi.org/10.1175/2009BAMS2618.1, 2009.

Bishop, C. H. and Hodyss, D.: Ensemble covariances adaptively localized with ECO-RAP. Part 2: A strategy for the atmosphere, Tellus, Series A: Dynamic Meteorology and Oceanography, 61 A, https://doi.org/10.1111/j.1600-0870.2008.00372.x, 2009.

Bishop, C. H. and Hodyss, D.: Adaptive ensemble covariance localization in ensemble 4D-VAR state estimation, Monthly Weather Review, 139, https://doi.org/10.1175/2010MWR3403.1, 2011.

Bishop, C. H., Etherton, B. J., and Majumdar, S. J.: Adaptive Sampling with the Ensemble Transform Kalman Filter. Part I: Theoretical Aspects, Monthly Weather Review, 129, 420–436, https://doi.org/10.1175/1520-0493(2001)129<0420:ASWTET>2.0.CO;2, 2001.

Bishop, C. H., Whitaker, J. S., and Lei, L.: Gain form of the ensemble transform Kalman Filter and its relevance to satellite data assimilation with model space ensemble covariance localization, Monthly Weather Review, 145, https://doi.org/10.1175/MWR-D-17-0102.1, 2017.

Burgers, G., Jan van Leeuwen, P., Evensen, G., Van Leeuwen, P. J., and Evensen, G.: Analysis scheme in the ensemble Kalman filter, Monthly Weather Review, 126, 1719–1724, https://doi.org/10.1175/1520-0493(1998)126<1719:ASITEK>2.0.CO;2, 1998.

Castruccio, F. S., Karspeck, A. R., Danabasoglu, G., Hendricks, J., Hoar, T., Collins, N., and Anderson, J. L.: An EnOI-Based Data Assimilation System With DART for a High-Resolution Version of the CESM2 Ocean Component, Journal of Advances in Modeling Earth Systems, 12, https://doi.org/10.1029/2020MS002176, 2020.

Chan, M.-Y.: Improving the analysis and forecast of tropical mesoscale convective systems through advancing the ensemble data assimilation of geostationary satellite infrared radiance observations, Ph.D. thesis, The Pennsylvania State University, State College, 2022.

Chan, M.-Y.: Code for PESE-GC Lorenz 96 study [Software], 2023.

Chan, M.-Y., Anderson, J. L., and Chen, X.: An efficient bi-Gaussian ensemble Kalman filter for satellite infrared radiance data assimilation, Monthly Weather Review, https://doi.org/10.1175/mwr-d-20-0142.1, 2020.

Chan, M.-Y., Chen, X., and Anderson, J. L.: The potential benefits of handling mixture statistics via a bi-Gaussian EnKF: tests with all-sky satellite infrared radiances, Journal of Advances in Modeling Earth Systems, https://doi.org/10.1029/2022MS003357, 2023.

Cheng, R. C. H. and Amin, N. A. K.: Estimating Parameters in Continuous Univariate Distributions with a Shifted Origin, Journal of the Royal Statistical Society: Series B (Methodological), 45, https://doi.org/10.1111/j.2517-6161.1983.tb01268.x, 1983.

El Gharamti, M.: Hybrid ensemble-variational filter: A spatially and temporally varying adaptive algorithm to estimate relative weighting, Monthly Weather Review, 149, https://doi.org/10.1175/MWR-D-20-0101.1, 2020.

Gaspari, G. and Cohn, S. E.: Construction of correlation functions in two and three dimensions, Quarterly Journal of the Royal Meteorological Society, 125, 723–757, https://doi.org/10.1256/smsqj.55416, 1999.

Gasperoni, N. A., Wang, X., and Wang, Y.: Using a Cost-Effective Approach to Increase Background Ensemble Member Size within the GSI-Based EnVar System for Improved Radar Analyses and Forecasts of Convective Systems, Monthly Weather Review, 150, https://doi.org/10.1175/mwr-d-21-0148.1, 2022.

Gordon, N. J., Salmond, D. J., and Smith, A. F.: Novel approach to nonlinear/non-gaussian Bayesian state estimation, IEE Proceedings, Part F: Radar and Signal Processing, 140, https://doi.org/10.1049/ip-f-2.1993.0015, 1993.

Grooms, I.: Analog ensemble data assimilation and a method for constructing analogs with variational autoencoders, Quarterly Journal of the Royal Meteorological Society, 147, https://doi.org/10.1002/qj.3910, 2021.

Grooms, I.: A comparison of nonlinear extensions to the ensemble Kalman filter: Gaussian anamorphosis and two-step ensemble filters, Computational Geosciences, 26, https://doi.org/10.1007/s10596-022-10141-x, 2022.

Huang, B. and Wang, X.: On the use of cost-effective Valid-Time-Shifting (VTS) method to increase ensemble size in the GFS hybrid 4DEnVar System, Monthly Weather Review, 146, https://doi.org/10.1175/MWR-D-18-0009.1, 2018.

Ide, K., Courtier, P., Ghil, M., and Lorenc, A. C.: Unified notation for data assimilation: Operational, sequential and variational, Journal of the Meteorological Society of Japan, https://doi.org/10.2151/jmsj1965.75.1B_181, 1997.

Kotsuki, S. and Bishop, C. H.: Implementing Hybrid Background Error Covariance into the LETKF with Attenuation-Based Localization: Experiments with a Simplified AGCM, Monthly Weather Review, 150, https://doi.org/10.1175/MWR-D-21-0174.1, 2022.

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

Kurosawa, K. and Poterjoy, J.: A Statistical Hypothesis Testing Strategy for Adaptively Blending Particle Filters and Ensemble Kalman Filters for Data Assimilation, Monthly Weather Review, 151, 105–125, https://doi.org/10.1175/MWR-D-22-0108.1, 2023.

Lei, L., Wang, Z., and Tan, Z. M.: Integrated Hybrid Data Assimilation for an Ensemble Kalman Filter, Monthly Weather Review, 149, https://doi.org/10.1175/MWR-D-21-0002.1, 2021.

Lorenz, E. N.: Predictability-a problem partly solved, in: Predictability of Weather and Climate, https://doi.org/10.1017/CBO9780511617652.004, 2006.

Monache, L. D., Anthony Eckel, F., Rife, D. L., Nagarajan, B., and Searight, K.: Probabilistic weather prediction with an analog ensemble, Monthly Weather Review, 141, https://doi.org/10.1175/MWR-D-12-00281.1, 2013.

Poterjoy, J.: A localized particle filter for high-dimensional nonlinear systems, Monthly Weather Review, 144, 59–76, https://doi.org/10.1175/MWR-D-15-0163.1, 2016.

Sakov, P., Oke, P. R., and Pav L S A Kov, B. E.: A deterministic formulation of the ensemble Kalman filter: an alternative to ensemble square root filters, https://doi.org/10.1111/j.1600-0870.2007.00299.x, 2008.

Skamarock, W., Klemp, J., Dudhi, J., Gill, D., Barker, D., Duda, M., Huang, X.-Y., Wang, W., and Powers, J.: A Description of the Advanced Research WRF Version 3. NCAR Tech. Note NCAR/TN–468+STR, 113 pp., NCAR TECHNICAL NOTE, https://doi.org/10.5065/D68S4MVH, 2008.

Snyder, C., Bengtsson, T., Bickel, P., and Anderson, J.: Obstacles to high-dimensional particle filtering, Monthly Weather Review, 136, https://doi.org/10.1175/2008MWR2529.1, 2008.

Sun, H., Lei, L., Liu, Z., Ning, L., and Tan, Z. M.: An Analog Offline EnKF for Paleoclimate Data Assimilation, Journal of Advances in Modeling Earth Systems, 14, https://doi.org/10.1029/2021MS002674, 2022.

Sun, H., Lei, L., Liu, Z., Ning, L., and Tan, Z. M.: A Hybrid Gain Analog Offline EnKF for Paleoclimate Data Assimilation, Journal of Advances in Modeling Earth Systems, 16, https://doi.org/10.1029/2022MS003414, 2024.

Tippett, M. K. and Delsole, T.: Constructed analogs and linear regression, Monthly Weather Review, 141, https://doi.org/10.1175/MWR-D-12-00223.1, 2013.

Tippett, M. K., Anderson, J. L., Bishop, C. H., Hamill, T. M., and Whitaker, J. S.: Ensemble Square Root Filters, Monthly Weather Review, 131, 1485–1490, https://doi.org/10.1175/1520-0493(2003)131<1485:ESRF>2.0.CO;2, 2003.

Van den Dool, H. M.: Searching for analogues, how long must we wait?, Tellus A, 46, https://doi.org/10.1034/j.1600-0870.1994.t01-2-00006.x, 1994.

van Leeuwen, P. J., Künsch, H. R., Nerger, L., Potthast, R., and Reich, S.: Particle filters for high-dimensional geoscience applications: A review, https://doi.org/10.1002/qj.3551, 2019.

Wan, E. and Van Der Merwe, R.: The unscented Kalman filter for nonlinear estimation, in: Proceedings of the IEEE 2000 Adaptive Systems for Signal Processing, Communications, and Control Symposium (Cat. No.00EX373), pp. 153–158, IEEE, https://doi.org/10.1109/ASSPCC.2000.882463, 2000.

Wang, X., Chipilski, H. G., Bishop, C. H., Satterfield, E., Baker, N., and Whitaker, J. S.: A multiscale local gain form ensemble transform Kalman filter (MLGETKF), Monthly Weather Review, 149, https://doi.org/10.1175/MWR-D-20-0290.1, 2021.

Whitaker, J. S. and Hamill, T. M.: Ensemble data assimilation without perturbed observations, Monthly Weather Review, 130, 1913–1924, https://doi.org/10.1175/1520-0493(2002)130<1913:EDAWPO>2.0.CO;2, 2002.

Whitaker, J. S. and Hamill, T. M.: Evaluating methods to account for system errors in ensemble data assimilation, Monthly Weather Review, 140, https://doi.org/10.1175/MWR-D-11-00276.1, 2012.

Xu, Q., Wei, L., Lu, H., Qiu, C., and Zhao, Q.: Time-expanded sampling for ensemble-based filters: Assimilation experiments with a shallow-water equation model, Journal of Geophysical Research Atmospheres, 113, https://doi.org/10.1029/2007JD008624, 2008.

Yuan, H., Lu, C., McGinley, J. A., Schultz, P. J., Jamison, B. D., Wharton, L., and Anderson, C. J.: Evaluation of short-range quantitative precipitation forecasts from a time-lagged multimodel ensemble, Weather and Forecasting, 24, https://doi.org/10.1175/2008WAF2007053.1, 2009.