A statistical hypothesis testing strategy for adaptively blending particle filters and ensemble Kalman filters for data assimilation

Kenta Kurosawa[a] Jonathan Poterjoy[a,b]

[a] *Department of Atmospheric and Oceanic Science, University of Maryland, College Park, Maryland*

[b] *NOAA Atlantic Oceanographic and Meteorological Laboratory, Miami, Florida*

Corresponding author: Kenta Kurosawa, kkurosaw@umd.edu

Early Online Release: This preliminary version has been accepted for publication in *Monthly Weather Review*, may be fully cited, and has been assigned DOI 10.1175/MWR-D-22-0108.1. The final typeset copyedited article will replace the EOR at the above DOI when it is published.

ABSTRACT: Particle filters avoid parametric estimates for Bayesian posterior densities, which alleviates Gaussian assumptions in nonlinear regimes. These methods, however, are more sensitive to sampling errors than Gaussian-based techniques such as ensemble Kalman filters. A recent study by the authors introduced an iterative strategy for particle filters that match posterior moments– where iterations improve the filter's ability to draw samples from non-Gaussian posterior densities. The iterations follow from a factorization of particle weights, providing a natural framework for combining particle filters with alternative filters to mitigate the impact of sampling errors. The current study introduces a novel approach to forming an adaptive hybrid data assimilation methodology, exploiting the theoretical strengths of non-parametric and parametric filters. At each data assimilation cycle, the iterative particle filter performs a sequence of updates while the prior sample distribution is non-Gaussian, then an ensemble Kalman filter provides the final adjustment when Gaussian distributions for marginal quantities are detected. The method employs the Shapiro-Wilk test to determine when to make the transition between filter algorithms, which has outstanding power for detecting departures from normality. Experiments using low-dimensional models demonstrate that the approach has significant value, especially for non-homogeneous observation networks and unknown model process errors. Moreover, hybrid factors are extended to consider marginals of more than one co-located variables using a test for multivariate normality. Findings from this study motivate the use of the proposed method for geophysical problems characterized by diverse observation networks and various dynamic instabilities, such as numerical weather prediction models.

Accepted for publication in *Monthly Weather Review.* DOI 10.1175/MWR-D-22-0108.1.

SIGNIFICANCE STATEMENT: Data assimilation statistically processes observation errors and model forecast errors to provide optimal initial conditions for the forecast, playing a critical role in numerical weather forecasting. The ensemble Kalman filter, which has been widely adopted and developed in many operational centers, assumes Gaussianity of the prior distribution and solves a linear system of equations, leading to bias in strong nonlinear regimes. On the other hand, particle filters avoid many of those assumptions but are sensitive to sampling errors and are computationally expensive. We propose an adaptive hybrid strategy that combines their advantages and minimizes the disadvantages of the two methods. The hybrid particle filter - ensemble Kalman filter is achieved with the Shapiro-Wilk test to detect the Gaussianity of the ensemble members and determine the timing of the transition between these filter updates. Demonstrations in this study show that the proposed method is advantageous when observations are heterogeneous and when the model has an unknown bias. Furthermore, by extending the statistical hypothesis test to the test for multivariate normality, we consider marginals of more than one co-located variable. These results encourage further testing for real geophysical problems characterized by various dynamic instabilities, such as real numerical weather prediction models.

1. Introduction

For convection-permitting numerical weather prediction systems, assimilating remotely-sensed observation networks (e.g., radar and cloudy radiance measurements) is required to depict mesoscale weather features accurately (e.g., Vukicevic et al. 2004; Stengel et al. 2009; Privé et al. 2013). It is well known that strongly nonlinear model dynamics and observation operators, however, can induce bias in Gaussian-based data assimilation methods that are commonly used for numerical weather prediction (e.g., Bocquet et al. 2010). Since ensemble Kalman filters (EnKFs; Evensen 1994; Houtekamer and Mitchell 1998; Evensen and van Leeuwen 2000) approximate prior densities using a Gaussian and solve a linear system of equations to adjust a sample of model states to fit the posterior mean and covariance, strongly nonlinear model dynamics and measurement operators can lead to bias, which impedes achieving accurate convection-permitting initial conditions for next-generation weather forecast models. This limitation is apparent for multi-scale weather prediction systems that exhibit large uncertainty in smaller scales, or when observations are sensitive to cloud processes (e.g., Poterjoy et al. 2017; Poterjoy 2022a). As a result, most

Accepted for publication in *Monthly Weather Review.* DOI 10.1175/MWR-D-22-0108.1.

infrared satellite data assimilation studies mainly focus on clear-sky observations (e.g., Errico et al. 2007; Fabry and Sun 2010; Minamide and Zhang 2017; Honda and Coauthors 2018). Therefore, developing new data assimilation methods that mitigate Gaussian assumptions is an active area of research.

One strategy, which has gained momentum in recent years, is to apply dimension-reduction procedures (namely, localization) to particle filters (PFs; Penny and Miyoshi 2016; Poterjoy and Anderson 2016; Poterjoy et al. 2017, 2019; Potthast et al. 2019). PFs avoid the parametric estimation of Bayesian posterior densities, thus providing great flexibility for solving a range of complex data assimilation problems. These methods, however, are more sensitive to sampling errors than EnKFs. As such, computational limitations pose a major obstacle, which has limited research examining the potential of PFs for operational weather prediction. Incorporating statistics from a large number of high-resolution ensemble members into the data assimilation step is one of the most effective ways to mitigate the effects of sampling errors, but this strategy is not often tractable.

Given the challenges discussed above, a natural progression is to combine PFs with methods that rely on parametric density estimates when appropriate. Several papers have proposed to hybridize PFs with EnKFs (Stordal et al. 2011; Frei and Kunsch 2013; Slivinski et al. 2015; Chustagulprom et al. 2016; Grooms and Robinson 2021) and with variational methods (Morzfeld et al. 2018). These methods are remarkabley accurate for cases of "moderate nonlinearity," which are characteristic of situations with a non-Gaussian priors but Gaussian-like posterior distribution (Metref et al. 2014; Morzfeld and Hodyss 2019; Grooms and Robinson 2021). For example, Frei and Kunsch (2013) introduced a procedure that makes a continuous transition between the ensemble and the particle filter update by factoring the likelihood. They choose a "splitting factor" to ensure that an effective ensemble size is maintained within a certain tolerance of a user-specified threshold. Based on this approach, Chustagulprom et al. (2016) developed a method to hybridize the general linear ensemble transform filter (LETF) framework and PFs, which can use observation-space localization and avoid linear assumptions for observation operators. Grooms and Robinson (2021) also introduced a filter that combines PFs with EnKFs, which is generally similar to Frei and Kunsch (2013) and Chustagulprom et al. (2016) in that it factors the likelihood. This method, like others, is effective for problems characterized by medium nonlinearity in model dynamics or

Accepted for publication in *Monthly Weather Review.* DOI 10.1175/MWR-D-22-0108.1.

measurement operators. In these papers, the value of the splitting factor is determined adaptively by the effective ensemble size. This choice is a heuristic one, thus motivating additional research into how to optimally combine PFs with EnKFs. For example, Nerger (2022) propose a method for estimating hybrid coefficients that is based not only the effective ensemble size but also the kurtosis and skewness of the ensemble. This method still requires the tuning of hyper-parameters, but generates more accurate filter estimates than using effective ensemble size alone. As mentioned in Chustagulprom et al. (2016), a more powerful and computationally feasible alternative is to adopt Kullback-Leibler divergence (KL divergence, Kullback and Leibler 1951) as a means of identifying proper choices of prior error distribution. This approach is one of the most frequently used objective functions to measure deviations from Gaussianity in forecast error distribution for weather models (Kondo and Miyoshi 2019; Li et al. 2019; Ruiz et al. 2021; Pimentel and Qranfal 2021). However, it is difficult to measure non-Gaussianity by the KL divergence for high-dimensional systems when the ensemble size is small or when a strange attractor makes numerical convergence and proper definition of the continuous limit complicated (Bocquet et al. 2010).

In this study, we introduce a novel approach to forming an adaptive PF-EnKF data assimilation methodology, which exploits the theoretical strengths of non-parametric (PF) and parametric (EnKF) filters. For this purpose, we use a recently proposed PF (Poterjoy 2022a,b), which introduces an iterative strategy for PFs that match posterior moments. For this PF, iterations improve the filter's ability to draw samples from non-Gaussian posterior densities despite fitting a limited number of moments. The iterations follow from a factorization of particle weights, which also provide a straightforward means of combining PFs with EnKFs to reduce the impact of sampling errors. To achieve the adaptive mixed methodology at each data assimilation cycle, we repeat the iterative PF update while the prior sample distribution is non-Gaussian, and update with EnKF when Gaussian distributions for marginal quantities are detected. Here, we introduce a statistical hypothesis testing approach to determine when to make this transition between filter algorithms. Several papers on data assimilation use statistical hypothesis tests for normality to measure the difference between the prior or posterior distribution and the normal distribution (e.g., Bocquet et al. 2010; Poterjoy 2016). However, research of incorporating a normality test directly into assimilation processes is unexplored.

Accepted for publication in *Monthly Weather Review.* DOI 10.1175/MWR-D-22-0108.1.

The current study first compares the power of several hypothesis tests by performing Monte Carlo simulations of data generated from choice distributions that are often used to characterize errors. Then we examine the newly developed mixed hybrid methodology employing the Shapiro-Wilk test (Shapiro and Wilk 1965), which has outstanding power among omnibus tests for detecting departures from normality (e.g., Srivastava and Hui 1987; Mendes and Pala 2003; Farrell et al. 2007; Villaseñor and González-Estrada 2009). Using the hypothesis testing approach allows us to accurately detect Gaussianity, even with small ensemble sizes, and explore possibilities other than effective ensemble size and the KL divergence for determining the splitting factor adaptively.

The manuscript is organized in the following manner. In Section 2, we briefly review the four well-known normality tests and compare the power of the tests. Section 3 introduces a statistical hypothesis testing approach to forming an adaptive PF-EnKF hybrid. We discuss the results and findings of numerical experiments conducted using low-dimensional toy models in Section 4. The last section discusses major findings from this study.

2. Power Comparisons of Four normality Tests

The assumption of a normal distribution is often an underlying premise of many academic fields and studies, including data assimilation. When the assumption of normality is violated, interpretations and inferences may lack reliability and validity. There are three commonly used procedures for evaluating whether a random independent sample comes from a normal population: graphical methods (histograms, box plots, Q-Q-plots), moment-based methods (skewness and kurtosis), and formal normality tests. In particular, a significant amount of normality tests have been proposed, and several studies have already compared their power (e.g., Dufour et al. 1998; Thadewald and Büning 2007; Razali and Wah 2011; Saculinggan and Balase 2013). In this section, we compare the power of four well-known formal tests of normality: Shapiro-Wilk test (SWT; Shapiro and Wilk 1965), Kolmogorov-Smirnov test (Kolmogorov 1933; Smirnov 1939), Lilliefors test (Lilliefors 1967), and Anderson-Darling test (Anderson and Darling 1954). The following sub-sections briefly review the four normality tests, and describe the simulation procedure.

Accepted for publication in *Monthly Weather Review.* DOI 10.1175/MWR-D-22-0108.1.

a. Methodology for the four hypothesis tests

1) Shapiro-Wilk Test

The normality test introduced by Shapiro and Wilk (1965) is the first test to detect departures from normality with skewness, kurtosis, or both (Althouse et al. 1998). It has become the most potent omnibus test in most situations because of its good power properties compared to a wide range of alternative tests. The basic idea behind SWT is to measure the goodness of fit of a straight line to a normal Q-Q plot (linear regression). Given an ordered random sample, $x_1 < x_2 < \ldots < x_n$, the original SWT statistic is defined as,

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_i\right)^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}, \tag{1}$$

where $\bar{x}$ is the sample mean. a_i is the expected values of the order statistics of independent and identically distributed random variables sampled from the standard normal distribution:

$$a_i = (a_1, \cdots, a_n) = \frac{\boldsymbol{m}^T \mathbf{V}^{-1}}{\left(\boldsymbol{m}^T \mathbf{V}^{-1} \mathbf{V}^{-1} \boldsymbol{m}\right)^{1/2}}, \tag{2}$$

where $\boldsymbol{m} = (m_1, \cdots, m_n)^T$. Here, the vector $\mathbf{m}$ consists of the expected values of the order statistics of independent random variables with identical distributions that are sampled from a normal distribution, and $\mathbf{V}$ is the covariance matrix of those order statistics.

The null hypothesis of SWT is that the data originate from a normally distributed population. Small values of W lead to the rejection of the null hypothesis, where $0 \leq W \leq 1$. The original SWT was limited to a sample size of 50 or less. Royston (1982) extended SWT to large samples and provided an approximation of the test statistic W, and Royston (1983) suggested a test for multivariate normality based on SWT. Royston (1992) arrived at an improved approximation to the weights, which allows SWT to effectively detect departures from multivariate normality for smaller sample sizes. Lastly, Royston (1995) introduced the FORTRAN algorithm AS R94, which is used in the current study. The algorithm includes a scaling process that sets the mean of the sample to zero and a centering process that sets the variance of the sample to one. Therefore, the algorithm allows the results to generalize not only to the standard normal distribution, but also to a normal distribution with a nonzero mean and non-unit variance.

Accepted for publication in *Monthly Weather Review.* DOI 10.1175/MWR-D-22-0108.1.

2) Kolmogorov-Smirnov Test

Kolmogorov-Smirnov test was first proposed by Kolmogorov (1933) and then improved by Smirnov (1939). The one-sample Kolmogorov-Smirnov test is a nonparametric test of the null hypothesis that the population cumulative distribution function (cdf) of the data is equal to the hypothesized cdf. Given an ordered random sample, the statistic is defined as,

$$D = \max_x(|F^*(x) - F_n(x)|), \tag{3}$$

where $F^*(x)$ is the cdf of the hypothesized distribution, and $F_n(x)$ is the empirical cdf. When the statistic value D is significant, the hypothesis that the sample comes from a normally distributed population is rejected.

3) Lilliefors Test

Kolmogorov-Smirnov test is appropriate when the hypothesized distribution parameters are completely known because the null distribution must be completely specified. In contrast, Lilliefors test, which is a modification of Kolmogorov-Smirnov test introduced by Lilliefors (1967), is a goodness-of-fit test for situations where the parameters of the null distribution are unknown and have to be estimated. Given an ordered random sample, the Lilliefors test statistic is defined as,

$$D = \max_x(|F^*(x) - S_n(x)|), \tag{4}$$

where $F^*(x)$ is the cdf of the hypothesized distribution, and $S_n(x)$ is the empirical cdf. The Lilliefors test statistic is the same as the Kolmogorov-Smirnov test statistic, but the tables of critical values of the two tests are different, leading to different conclusions and decisions. In Kolmogorov-Smirnov test, we must completely give the null distribution. On the other hand, Lilliefors test is a two-sided goodness-of-fit test and is powerful when the parameters of the null distribution are unknown.

Accepted for publication in *Monthly Weather Review.* DOI 10.1175/MWR-D-22-0108.1.

4) Anderson-Darling Test

Anderson-Darling test, introduced by Anderson and Darling (1954), is a modification of the Cramer-von Miles test (Cramér 1928). The Anderson-Darling test statistic is defined as,

$$n \int_{-\infty}^{\infty} (F_n(x) - F(x))^2 \, w(x) dF(x), \tag{5}$$

where n is the sample size, $w(x)$ is a weight function, $F(x)$ is the hypothesized distribution, and $F_n(x)$ is the empirical cdf. The weight function is defined as

$$w(x) = [F(x)(1 - F(x))]^{-1}, \tag{6}$$

and Arshad et al. (2003) suggested the following formula as the test statistic of Anderson-Darling test:

$$A_n^2 = -n - \sum_{i=1}^{n} \frac{2i-1}{n} \left[\ln\left(F\left(X_i\right)\right) + \ln\left(1 - F\left(X_{n+1-i}\right)\right) \right], \tag{7}$$

where $X_1 < X_2 < \ldots < X_n$ are the ordered sample data points. The weighting function (6) is more sensitive to outliers because the weights of the observed values at the edges of the distribution are larger. Therefore, it is especially suited for detecting deviations from normality at the tails of the distribution.

b. Simulation Procedures

Monte Carlo simulations are the most commonly used approaches to compare and evaluate the accuracy of a hypothesis test in detecting the degree of contamination by outliers and the test power with respect to sample size. Following previous studies, we use Monte Carlo simulations to evaluate the power of SWT, Kolmogorov-Smirnov test, Lilliefors test, and Anderson-Darling test statistics in testing whether a random sample of n independent observations is obtained from a population with a normal $N(\mu, \sigma^2)$ distribution. Several papers have already shown the superiority of SWT over the other tests (e.g., Mendes and Pala 2003; Razali and Wah 2011). Thus, the simulation in this section focuses on a larger selection of distributions than previous studies, which are motivated by the diverse shapes of error distributions found for geophysical models and observing systems, and aims to reconfirm the superiority of SWT over the other tests.

Accepted for publication in *Monthly Weather Review.* DOI 10.1175/MWR-D-22-0108.1.

As summerized in Table 1, we examine the following five cases of nine distributions to cover a variety of standardized skewness ($\sqrt{\beta_1}$) and kurtosis (β_2) : $N(0,1)$, $U(0,1)$, $Beta(2,2)$, $Logistic(0,1)$, $t(5)$, $Weibull(3,5)$, $Beta(2,4)$,$Gamma(0,1)$, and $\chi^2(10)$. The values of $\sqrt{\beta_1}$ and β_2 for each distribution are summarized in Table 1. For each distribution, we set the significance level at 0.05, the sample sizes at n = 5, 20, 40, 100, 300, 500, 1000, and the number of trials at 100,000. The null and alternative hypotheses of the four tests are:

H_0: The distribution is normal

H_1: The distribution is not normal

As summarized in Table 2, the "test power" (true positive) of a hypothesis test is the probability that the test correctly rejects the null hypothesis when the alternative hypothesis is true. On the other hand, the "type I error" (false positive) is the error of rejecting the null hypothesis when it is actually true. Therefore, if a sample is taken from $N(0,1)$ population, the number of rejected H_0 hypotheses is the probability of a type I error (Case A in Table 1). In contrast, if the samples are from a population that is not normal distribution, the number of H_0 rejected is the power of the test (Cases B-E in Table 1).

In Table 2, the probability of a type I error occurring is denoted by α and the probability of a type II error by β. The probability of both the type I and type II errors should be small. However, it is impossible to make both small because the risk ratios α and β are in a trade-off relationship. Generally speaking (in society), committing a type I error is often a more serious problem. Moreover, as discussed in detail in Section 3, in the current study, we prefer erring on the side of a lower type I error in order to avoid the use of an EnKF when the distribution is truly non-Gaussian. This is because the PF update can provide adequate estimates for both non-Gaussian or Gaussian prior distributions; the same cannot be said about the EnKF. Therefore, the correct procedure for hypothesis testing is to determine the acceptable risk rate α in advance, and then select the hypothesis test method with the highest test power $1-\beta$ among them. Hence, in the current study, we focus on the type I error for $N(0,1)$ and on the test power for the other distributions (Cases B-E in Table 1); We will not look at the type II error and specificity in this study.

Accepted for publication in *Monthly Weather Review.* DOI 10.1175/MWR-D-22-0108.1.

TABLE 1. Classification of cases by skewness and kurtosis of the distribution

Case	skewness ($\sqrt{\beta_1}$), kurtosis (β_2)	Distributions
A	$\sqrt{\beta_1} = 0, \beta_2 = 3$	$N(0,1)$
B	$\sqrt{\beta_1} = 0, \beta_2 < 3$	$U(0,1)$, $Beta(2,2)$
C	$\sqrt{\beta_1} = 0, \beta_2 > 3$	$Logistic(0,1)$, $t(5)$
D	$\sqrt{\beta_1} \neq 0, \beta_2 < 3$	$Weibull(3,5)$, $Beta(2,4)$
E	$\sqrt{\beta_1} \neq 0, \beta_2 > 3$	$Gamma(2,2)$, $\chi^2(10)$

TABLE 2. Definitions of terminologies in a statistical test

		Decision	
		Accept H_0	Reject H_0
Null Hypothesis H_0 is	True	**Specificity** "True Negative" Probability: $1-\alpha$	**Type I Error** "False Positive" Probability: α
	False	**Type II Error** "False Negative" Probability: β	**Test Power** "True Positive" Probability: $1-\beta$

c. Results

Figure 1 shows the variation of the type I error (Fig.1a) and test power (Fig.1b-i) with the sample sizes n for the four tests for each distribution when $\alpha = 0.05$. When the distribution is $N(0,1)$, all four tests generally achieved $\alpha = 0.05$. In the case of the symmetric distributions ($\sqrt{\beta_1} = 0$; Fig.1b-e), SWT is the best, followed by Anderson-Darling test, Lilliefors test, and Kolmogorov-Smirnov test. However, all tests have low power when the sample size is less than 100. In particular, when β_2 is greater than 3, the power of Kolmogorov-Smirnov test is significantly inferior to the other three (Fig.1d,e). All other tests can attain 80% power when the sample size is 1000. The power for the asymmetric distributions ($\sqrt{\beta_1} \neq 0$; Fig.1f-i) is also highest for SWT, followed by Anderson-Darling test, Lilliefors test, and Kolmogorov-Smirnov test. For the $Weibull(3,5)$ distribution, the overall power is low because the kurtosis β_2 is close to 3 (Fig.1f). In the other cases, however, (Fig.1g-i), SWT and Anderson-Darling test require 200 samples to achieve 90% power, while Kolmogorov-Smirnov test requires 500 samples; Lilliefors test shows power for sample sizes between these two numbers.

Accepted for publication in *Monthly Weather Review.* DOI 10.1175/MWR-D-22-0108.1.

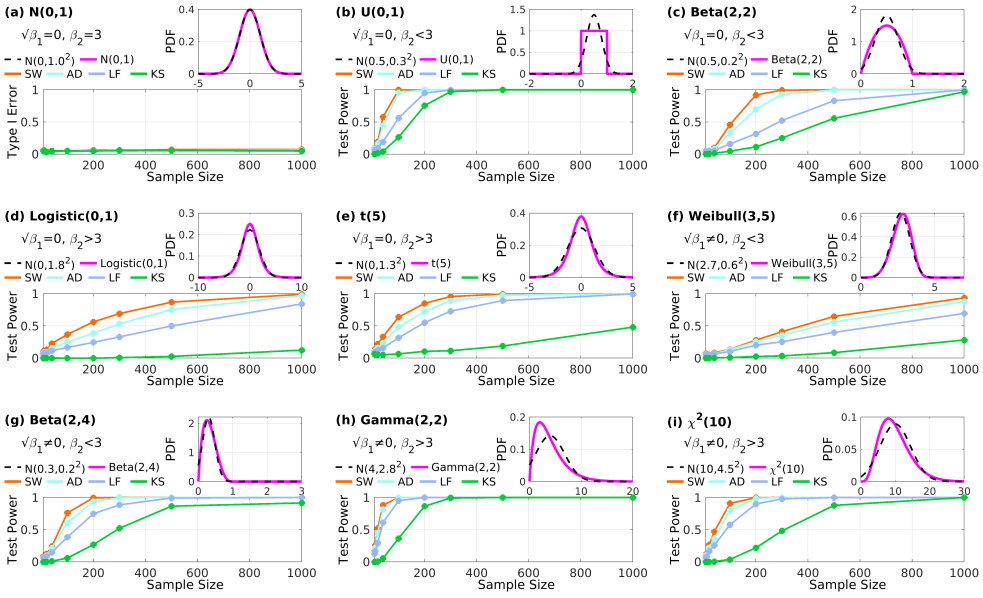

FIG. 1. Type I Error (a) and test power (b-i) of Shapiro-Wilk test (SW; orange), Anderson-Darling test (AD; cyan), Lilliefors test (LF; blue), and Kolmogorov-Smirnov test (KS; green) for different distributions and sample sizes. The magenta line shows each distribution, and the black dash line shows a close normal distribution to each distribution. The distributions cover a variety of standardized skewness ($\sqrt{\beta_1}$) and kurtosis (β_2).

The overall results show that in all cases the power of SWT is superior to the other tests for small sample sizes, which is the regime of interest for ensemble data assimilation applied to weather models. This result is generally consistent with the results of many other previous studies mentioned before. Therefore, this paper uses SWT to detect Gaussianity in the new hybrid method hereafter. Note that "failing to reject the null hypothesis" is not the same as "accepting the null hypothesis." In such a case, it is still not exactly clear whether the null or alternative hypothesis is correct. For simplicity, this study interprets the null hypothesis to be that "the samples are from a population that follows a normal distribution." If the null hypothesis of SWT is not rejected then the samples are assumed to come from a Gaussian distribution.

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-22-0108.1.

3. Implementation with the local particle filter

This section presents the mathematical framework for implementing the adaptive PF-EnKF hybrid method by embedding SWT, the most powerful statistical test in the previous section, into the local PF. For the purposes of the adaptive hybrid methodology, the current study uses the recently proposed PF by Poterjoy (2022b) that introduces an iterative strategy, which will be denoted P22 hereafter. We briefly describe the parts of PF that are relevant to the implementation of the proposed method.

The local PF operates by assimilating observations with independent errors sequentially and combining sampled particles and prior particles for each observation. By serially processing an observation y in a sequence of observations and updating particles after each observation space sampling step, posterior particles can be adjusted in a manner consistent with bootstrap sampling. The n^{th} updated particle $\mathbf{x}_y^n$ is represented by the linear combination of the re-sampled particle $\mathbf{x}^{k_n}$, conditioned on all observations before y, and the prior particle $\mathbf{x}^n$ as follows:

$$\mathbf{x}_y^n = \overline{\mathbf{x}}_y + \mathbf{r}_1 \circ (\mathbf{x}^{k_n} - \overline{\mathbf{x}}_y) + \mathbf{r}_2 \circ (\mathbf{x}^n - \overline{\mathbf{x}}_y), \tag{8}$$

where k_n is the index of each sampled particle, and $\overline{\mathbf{x}}_y$ is the localized posterior mean accumulating the full weight of all observations up to y. $\mathbf{r}_1$ and $\mathbf{r}_2$ are vectors of coefficients that ensure the update satisfies the posterior mean and variance of marginals everywhere in state space—as depicted by importance weights.

Poterjoy et al. (2019) introduced several filter stabilization strategies in the PF of Poterjoy (2016) to avoid particle degeneracy. In particular, regularization and tempering are effective methods when sampling errors are large, and the sample size is small. Regularization is equivalent to increasing the particle weights up to a power β by inflating the observation error variance. This regularization allows the particles to acquire a specific "effective sample size N_{eff}", and is particularly helpful in stabilizing the filter when all particles are far from an observation. The regularization provides temporary iterations for the local PF, which is a posterior tempering strategy. This iterative approach also improves the filter's ability to sample from non-Gaussian posterior densities, even though it fits a limited number of moments. The iterations consist of a factorization of particle weights, thus providing a natural framework for combining the local PF with alternative filters to reduce the

Accepted for publication in *Monthly Weather Review.* DOI 10.1175/MWR-D-22-0108.1.

impact of sampling error. When provided with Gaussian likelihoods, a partial update performed by a PF can adjust particles to more closely resemble samples from a Gaussian, even if the prior exhibits a complex nn-Gaussian shape. The resulting intermediate update then makes the EnKF an appropriate choice for the remaining update (Grooms and Robinson 2021). P22 introduced the hybrid parameter $\boldsymbol{\kappa}$, which is an N_x-dimensional vector that determines when to switch from a PF update to a parametric filter update. For the iterative PF, the hybrid parameter κ and "the target effective sample size N_{eff}^t" are required to be specified by users. PF updates are repeated until $\sum_{k=1}^{N_k} \beta_{j,k} = \boldsymbol{\kappa}_j$ for $0 \leq \boldsymbol{\kappa}_j \leq 1$ at the j^{th} grid point, where N_k is the number of iterations. Here, β_j enforces a minimum N_{eff} for weight at the j^{th} grid point. When N_{eff} is below N_{eff}^t, β is determined adaptively by eq. (28) in Poterjoy et al. (2019) so that N_{eff}^t is satisfied. Following the initial set of local PF iterations, the last adjustment is performed using an EnKF with the measurement error variance $\mathbf{R}$ inflated by the factor $\frac{1}{\boldsymbol{\eta}_j}$, where $\boldsymbol{\eta}_j = 1 - \boldsymbol{\kappa}_j$. For example, to hybridize the PF and EnKF in the ratio of 7:3 at the j^{th} grid point, $\boldsymbol{\kappa}_j$ is set to 0.7 in the first place, and PF updates are repeated until $\sum_{k=1}^{N_k} \beta_{j,k} = \boldsymbol{\kappa}_j = 0.7$. The value of $\beta_{j,k}$ is determined adaptively based on N_{eff}^t in the k^{th} iteration, which means N_k is determined adaptively as well and different at each grid point. Therefore, since N_k at j^{th} grid point is determined by $\beta_{j,k}$, N_k becomes larger when N_{eff}^t is set to a larger value, and vice versa. Lastly, an EnKF update is performed with $\mathbf{R}$ inflated by a remaining factor $\boldsymbol{\eta}_j$, where $\boldsymbol{\eta}_j = 1 - \boldsymbol{\kappa}_j = 0.3$. Note that, for simplicity, P22 uses the same value for $\boldsymbol{\kappa}_j$ at all grid points and in all data assimilation cycles.

In the current study, we allow $\boldsymbol{\kappa}$ and $\boldsymbol{\eta}$ to be adjusted adaptively through space and time during data assimilation, while these values are held constant and set heuristically through tuning in P22. We repeat the PF update until SWT suggests that particles are samples from a Gaussian; $\sum_{k=1}^{N_k} \beta_{j,k}$ up to this point is defined as $\boldsymbol{\kappa}_j$, and the value of $\boldsymbol{\eta}_j$ is determined by $1 - \boldsymbol{\kappa}_j$. Once $\boldsymbol{\kappa}_j$ at all grid points has been determined, we perform the serial ensemble square-root filter (serial EnSRF; Whitaker and Hamill 2002) with $\mathbf{R}$ inflated by the inverse of $\boldsymbol{\eta}_j$ as the final adjustment. Thus, the Kalman gain matrix at the j^{th} grid point when the i^{th} observation is assimilated is described as follows:

$$\mathbf{K}_j = \mathbf{E}_j^f \mathbf{D}_i^{f\top} \left(\mathbf{D}_i^f \mathbf{D}_i^{f\top} + \frac{1}{\boldsymbol{\eta}_j} \mathbf{R}_i \right)^{-1}, \tag{9}$$

where $\mathbf{E}^f$ consists of model-space forecast ensemble perturbations and $\mathbf{D}^f$ consists of observation-space forecast ensemble perturbations, with both matrices normalized by $\frac{1}{\sqrt{Ne-1}}$. Note that the use of a tangent linear measurement operator in (9) is avoided in the current study and most others by assimilating observations serially.

In the case that SWT does not detect Gaussianity during N_k iterations at the j^{th} grid point, $\boldsymbol{\kappa}_j$ and $\boldsymbol{\eta}_j$ become 1 and 0, respectively, and no EnKF update is performed at the grid point. Similarly, if SWT detects Gaussianity in the first iteration at the j^{th} grid point, then $\boldsymbol{\kappa}_j = 0$ and $\boldsymbol{\eta}_j = 1$, and no PF update is performed at the grid point. Thus, in situations where the posterior is clearly non-Gaussian, the filter can have the option of retaining the local PF. The hybrid approach aims to obtain an intermediate distribution that is closer to Gaussian than the prior distribution by the PF updates, and to make this intermediate distribution closer to Gaussian by the EnKF. In cases that we cannot obtain an intermediate distribution closer to Gaussian, we can perform the iterative PF updates alone, without using EnKF in the last step, which is the strength of the adaptive strategy. The advantage of using the EnKF in the last step if a Gaussian is encountered during iterations is purely due to it being a more robust choice when ensemble sizes are small (and the distribution is indeed Gaussian).

As in P22, the N^t_{eff} still needs to be specified by the user and this parameter can influence the results. In general, N^t_{eff} determines when filter updates are made during iterations. High N^t_{eff} typically leads to more iterations and a larger final effective ensemble size than a small N^t_{eff}. This choice is ultimately a trade-off between the frequency of performing SWT and cost of implementation.

Since the computational cost of SWT is not expensive, the adaptive approach introduced in the current study, which incorporates the statistical test into the local PF introduced by P22, is generally less computationally expensive than the iterative LPF. Under most circumstances, the hybrid requires fewer iterations, thus leading to a cost saving. Nevertheless, the PF introduced by P22 is computationally more costly than pure EnKF because of the use of regularization and tempering. For more information, please refer to Poterjoy (2022a) and P22.

Note that $\boldsymbol{\kappa}$ and $\boldsymbol{\eta}$ are uniquely specified for each observation-space prior variable as well. In this case, they are N_y-dimensional vectors and we again use SWT to determine when each element

Accepted for publication in *Monthly Weather Review.* DOI 10.1175/MWR-D-22-0108.1.

of observation-space forecast ensembles may follow a Gaussian distribution. The κ and η defined for the observation-space are used for the observation-space filter updates.

In summary, the adaptive hybrid PF-Serial EnSRF with SWT are realized by Algorithms 1 – 2. In both algorithms, $\mathbf{x}^f$ and $\mathbf{x}^a$ are N_x-dimensional background and analysis vectors, respectively, and $\mathbf{y^o}$ is an N_y-dimensional set of observations. H is a observation operator that maps a model state to its corresponding observation state:

$$\mathbf{y}^f = H_i(\mathbf{x}^f), \tag{10}$$

where H_i is the measurement operator for the i^{th} observation.

Accepted for publication in *Monthly Weather Review.* DOI 10.1175/MWR-D-22-0108.1.

Algorithm 1 Adaptive mixed PF-EnKF update with SWT

function PF_ENKF_HYBRID
  $k = 1$
  $\boldsymbol{\kappa} = \mathbf{0.0}$ ▷ Vector with N_x dimensions
  $\boldsymbol{\kappa}\_residual = \mathbf{1.0} - \boldsymbol{\kappa}$
  while $\max(\boldsymbol{\kappa}\_residual) > 0$ **do** ▷ Tempering
    for $j = 1{:}N_x$ **do**
      if $\boldsymbol{\kappa}\_residual(j) > 0$ **then**
        $SWT\_result \leftarrow \text{SWT}(\mathbf{x}_j^f)$ ▷ Shapiro-Wilk Test
        if SWT_result = $Gaussian$ **then**
          $\boldsymbol{\eta}_j = \boldsymbol{\kappa}\_residual(j)$
          $\boldsymbol{\kappa}\_residual(j) = \mathbf{0.0}$
        end if
      end if
    end for
    $(\beta_k, \boldsymbol{\kappa}\_residual) \leftarrow \text{Regularization}(\boldsymbol{\kappa}\_residual)$
    for $i = 1{:}N_y$ **do**
      $\mathbf{x}^a \leftarrow$ The Local PF $(\mathbf{x}^f, \mathbf{y}_i^o, \beta_k)$ ▷ The Local PF core
      $\mathbf{x}^f \leftarrow \mathbf{x}^a$
    end for
    $k = k + 1$
  end while
  $\mathbf{x}^a \leftarrow \text{EnKF\_tempered}(\mathbf{x}^f, \mathbf{y^o}, \boldsymbol{\eta})$ ▷ EnKF as the last adjustment
end function

Accepted for publication in *Monthly Weather Review.* DOI 10.1175/MWR-D-22-0108.1.

Algorithm 2 Serial EnSRF update with inflated $\mathbf{R}$

function EnKF_tempered($\mathbf{x}^f, \mathbf{y}^o, \boldsymbol{\eta}$)
  for $i = 1{:}N_y$ **do**
    $\mathbf{y}^f = H_i(\mathbf{x}^f)$
    $\mathbf{E}^f = \frac{1}{\sqrt{N_e - 1}}[\delta\mathbf{x}_1^f \mid \cdots \mid \delta\mathbf{x}_{N_e}^f]$
    $\mathbf{D}_i^f = \frac{1}{\sqrt{N_e - 1}}[\delta\mathbf{y}_1^f \mid \cdots \mid \delta\mathbf{y}_{N_e}^f]$
    for $j = 1{:}N_x$ **do**
      $\mathbf{K}_j = \mathbf{E}_j^f \mathbf{D}_i^{f\top} \left(\mathbf{D}_i^f \mathbf{D}_i^{f\top} + \frac{1}{\eta_j}\mathbf{R}_i\right)^{-1}$
    end for
    $\bar{\mathbf{x}}^a = \bar{\mathbf{x}}^f + \mathbf{K}(\mathbf{y}_i^o - \mathbf{y}^f)$
    $\alpha = (1 + \sqrt{\frac{\mathbf{R}_i}{\mathbf{D}_i^f \mathbf{D}_i^{f\top} + \mathbf{R}_i}})^{-1}$
    $\tilde{\mathbf{K}} = \alpha\mathbf{K}$
    $\mathbf{E}^a = \mathbf{E}^f - \tilde{\mathbf{K}}\mathbf{D}_i^f$
    $\mathbf{x}^a = \bar{\mathbf{x}}^a + \mathbf{E}^a$
    $\mathbf{x}^f \leftarrow \mathbf{x}^a$
  end for
end function

Accepted for publication in *Monthly Weather Review.* DOI 10.1175/MWR-D-22-0108.1.

4. Numerical experiments with low-order models

This section explores the behavior of the newly developed method through numerical simulations. In the first experiment, we use a simple univariate problem to illustrate the difference of the adaptive hybrid method between iterative EnKF and bootstrap PF using tempering. In the second experiment, we use the 40-variable dynamic model of Lorenz (1996) to compare the advantages of the adaptive method over EnKF, the local PF, and hybrid PF-EnKF with fixed values of κ and η. These experiments use simulated measurements to target several scenarios, such as varying spatial density, highly nonlinear dynamics, mixed measurement operators, and unresolved model error. The last experiment uses an idealized kinematic vortex, which was used in Poterjoy (2022a) to replicate findings from real-data applications. The kinematic vortex model allows us to generate observations that emulate realistic observations for an application containing large spatial dependence across variables, while retaining great flexibility in our construction of data assimilation experiments. Among the common parameters used for idealized data assimilation applications, such as observation error variance, observation locations, and ensemble size, these experiments contain parameters that indirectly control the shape of the full multivariate prior, thus allowing for an analysis of the adaptive hybrid technique under controllable conditions.

a. Univariate application

Using a univariate example, we can visualize how the newly proposed adaptive hybrid method works compared to filters that use iterative strategies. We compare three iterative filters in this section: EnKF with the multiple data assimilation scheme (EnKF-MDA) proposed by Emerick and Reynolds (2012), bootstrap filter adopting the iterative approach (IPF), and a hybrid of the IPF and EnKF (adaptive IPF-EnKF). The number of iteration is set to four for EnKF-MDA and IPF for this demonstration. For EnKF-MDA, when the same observation is assimilated N_a times, the inflated measurement error covariance matrix is used:

$$\mathbf{K} = \mathbf{E}^f \mathbf{D}^{f\top} (\mathbf{D}^f \mathbf{D}^{f\top} + \alpha_i \mathbf{R})^{-1}, \tag{11}$$

where

$$\sum_{i=1}^{N_a} \frac{1}{\alpha_i} = 1. \tag{12}$$

Accepted for publication in *Monthly Weather Review.* DOI 10.1175/MWR-D-22-0108.1.

In this experiment, we use $\alpha_i = 4$ for $i = 1, \ldots, N_a$, where $N_a = 4$. For further details on EnKF-MDA, we encourage readers to review the mathematical descriptions in Emerick and Reynolds (2012). The IPF also uses a factorization of the likelihood to break the PF update step into a sequence of four updates, namely $\beta_i = \frac{1}{4}$ for $i = 1, \ldots, N_a$, where $N_a = 4$. For the adaptive IPF-EnKF, we first set $\beta_i = \frac{1}{8}$ and repeat the bootstrap PF update until SWT detects that prior members are samples from a Gaussian. We then replace the remaining PF update with an EnKF update using R inflated with the inverse of the remaining likelihood ($\frac{1}{1-\kappa} = \frac{1}{\eta}$). Note that the multiple updates in EnKF-MDA and IPF are identical to single updates of each because the operator is linear in this example.

Consider the example shown in Fig. 2, where 10^4 prior members are updated using an observation whose value is 4 and observation error standard deviation is set to $\sigma_y = 0.8$. Among the 10^4 prior members, three quarters are selected from $N(-4, 1.2^2)$, while the rest are from $N(3.5, 1.2^2)$, whose mean value is smaller than the observation. Therefore, the prior ensemble is a bimodal distribution. Using the same prior for all three filters, Figs. 2a-c show the posterior distribution after the first iteration. In all cases, we can see that each filter shifts the ensemble toward the observation, however the EnKF-MDA inherits the bimodal distribution of the prior for the posterior distribution, while the IPF and IPF-EnKF correctly retain a single mode[1]. The bimodal posterior distributions in the EnKF-MDA are not relieved by the completion of all iterations (Fig. 2j). In the IPF, after all the updates, the posterior pdf is relatively close to the likelihood of the observation, but exhibits negative skewness because numerous particles remain in the left-most mode (Fig. 2k). In the adaptive IPF-EnKF case, SWT detected Gaussianity in the distribution of the ensembles after three iterations of the bootstrap PF (Fig. 2i), and then EnKF was performed using R inflated by $\frac{8}{5}$, which is the inverse of the remaining observation error variance (Fig. 2l). As a result, the IPF-EnKF posterior is close to the IPF, indicating that the hybrid method correctly transitioned to the partial EnKF step once a Gaussian distribution was detected. Furthermore, we emphasize that the univariate application is presented for illustration only, as the IPF-EnKF is not expected to provide benefits over the IPF when the ensemble size is large.

[1]For the provided ensemble size, the last iteration of the IPF is an accurate estimate of the true Bayesian posterior

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-22-0108.1.

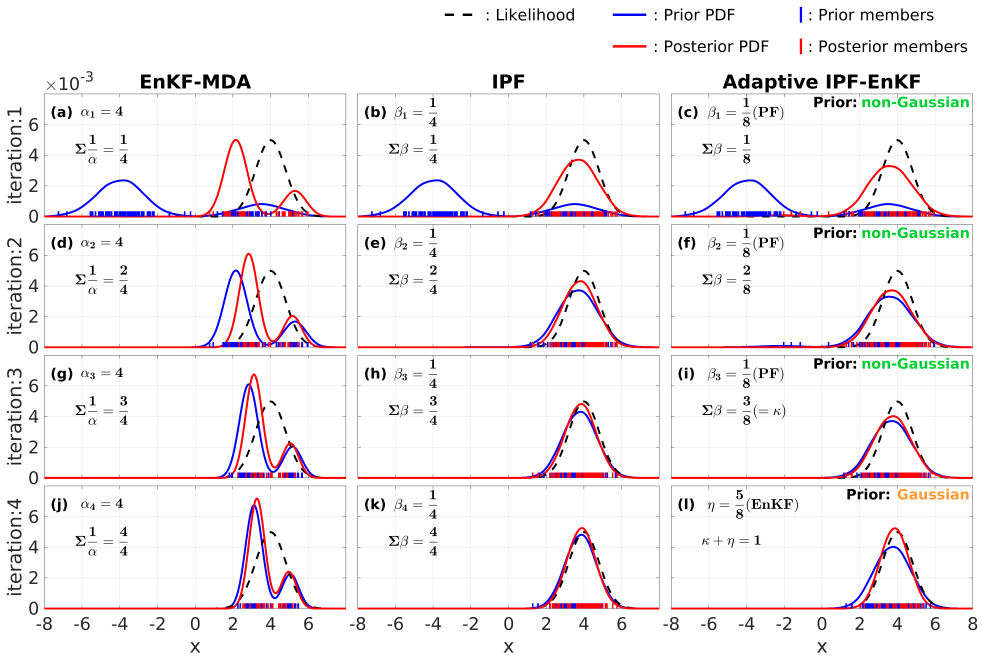

FIG. 2. A univariate example of how the updates differ in each iterative filter for EnKF-MDA (first column), IPF (second column), and adaptive mixed IPF-EnKF (third column). Each row corresponds to an iteration. The blue and red lines indicate marginal prior and posterior pdfs, respectively. The black dashed line indicates the observation likelihood.

b. 40-variable dynamical system

1) EXPERIMENTAL DESIGNS

For the next set of experiments, we assess the proposed adaptive hybrid strategy through idealized numerical experiments with the Lorenz 40-variable model (Lorenz 1996; Lorenz and Emanuel 1998), denoted L96 hereafter. The model consists of N_x equally-spaced variables and is defined by

$$\frac{dx_i}{dt} = (x_{i+1} - x_{i-2})x_{i-1} - x_i + F, \tag{13}$$

where $i = 1, 2, \ldots, N_x$ with cyclic boundaries: $x_{i+N_x} = x_i$ and $x_{i-N_x} = x_i$. The model is integrated forward numerically using the fourth-order Runge-Kutta scheme and a model time step of 0.05 non-dimensional units, which is corresponding to 6h (Lorenz 1996). As in Lorenz (1996), we

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-22-0108.1.

fix N_x at 40 and use $F = 8.0$, except for one set of experiments that consider an imperfect model case; in this case, measurements are simulated from a model trajectory with $F = 8.0$, but the model forcing F is fixed at 9.0.

Experiments include three forms of measurement operator H as in Kurosawa and Poterjoy (2021): "Linear Case", "Nonlinear Case 1", and "Nonlinear Case 2" use $H(\mathbf{x}) = \hat{\mathbf{x}}$, $H(\mathbf{x}) = \hat{\mathbf{x}} \circ \hat{\mathbf{x}}$, and $H(\mathbf{x}) = \log[ABS(\hat{\mathbf{x}})]$, respectively. Here, $\hat{\mathbf{x}}$ is a subset of N_y variables in $\mathbf{x}$ chosen by H, and ABS stands for the absolute value of each element. Uncorrelated Gaussian errors selected from $N(0, \sigma_y^2 I)$ are added to each operator: $\sigma_y = 1.0$ for the first two experiments, while $\sigma_y = 0.1$ for the third case because of the smaller information content provided by this observation. All experiments use $N_y = 20$ observations applying one or two of the operators. When only one observation operator is used, there are three settings: a setting with evenly distributed observations and $F = 8.0$ ("normal"), a setting with evenly distributed but $F = 9.0$ ("model error"), and a setting with missing observations in some places ("data void"). In contrast to the setting where observations are homogeneous throughout the domain, the "data void" setting is designed to target the heterogeneous observation network of real atmospheric models. Note that for this setting, we set the observation points at grid points $1-10, 21-30$. In the "mix" case, two observation operators are used, namely the first and the second half of the observation points use different observation operators. The experimental settings for each of these cases are summarized in Table 3a.

All experiments in this section use an observation frequency of 6 h. Observations are assimilated over a 10-year period, and root-mean-square errors (RMSEs) from the last 9 years are used to quantify the accuracy of the posterior analyses, ignoring the first year spin up period. In this set of experiments, we perform 100 parallel trials out of an abundance of caution. For localization, we use the fifth-order correlation function controlled by a radius of influence (ROI) given by Gaspari and Cohn (1999). For posterior inflation, the current study adopts the strategy known as "relaxation to prior perturbation" (RTPP; Zhang et al. 2004) after EnKF update. Similar to the α used in the relaxation method, for the local PF, we use a mixing parameter γ to maintain particle diversity during updates in (8). When the ensemble size is small, this parameter works to prevent filter divergence. γ is a scalar between 0 and 1, and acts to increase diversity in particles without modifying prior or posterior error variance. Each time the particles in state space are updated, the prior particles are mixed with the re-sampled particles (Poterjoy 2022b). The target N_{eff} is

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-22-0108.1.

TABLE 3. Configuration of cycling data assimilation experiments.

(a)

Expt	$H(x)$	F	Observation Points
Linear Case	x	8.0 | 9.0	evenly distributed | data void
Nonlinear Case 1	x^2	8.0 | 9.0	evenly distributed | data void
Nonlinear Case 2	$\log(\|x\|)$	8.0 | 9.0	evenly distributed | data void
Mix Case 1	x & x^2	8.0	evenly distributed
Mix Case 2	x & $\log(\|x\|)$	8.0	evenly distributed
Mix Case 3	x^2 & $\log(\|x\|)$	8.0	evenly distributed

(b)

N_e	ROI	α,γ
10	2 | 1	0.3
20	5 | 1	0.3
40	7 | 2	0.3
100	9 | 3	0.3
300	9 | 3	0.3

fixed at $N_{\text{eff}}^t = 0.5 \times N_e$ for all experiments. We configure the number of members $10, 20, 40, 100$, and 300, and arbitrarily tune all filter parameters, namely ROI, α, and γ, for each ensemble size. For the "normal" and "mix" settings, ROIs for $N_e = 10, 20, 40, 100$, and 300 are $2, 5, 7, 9$, and 9, respectively. On the other hand, to ensure the stability of the experiments, for the "data void" and "model error" settings, ROIs for $N_e = 10, 20, 40, 100$, and 300 are $1, 1, 2, 3$, and 3, respectively. The settings of filter parameters are summarized in Table 3b. Under each experimental setting, we performed a total of 12 experiments, one in which the value of κ is estimated adaptively, and the others in which κ is fixed at 0.1 increments from 0 to 1. Note that, in this section, the inflation and localization parameters for EnKF, the local PF, and hybrid experiments are unified, so we limited this tuning to experiments that use the LPF and EnKF alone. The tuning step is complicated for hybrid implementations, since we would have different optimal values for ROI and other parameters as soon as we change the κ value. This feature makes it difficult to identify optimal parameters in a cost-effective manner. While we acknowledge this limitation in the comparisons, we note that hybrid configurations still tend to outperform the LPF and EnKF despite not following a rigorous tuning. In other words, we believe that the use of optimal parameters may slightly change the results of the following experiments, but it will not change the conclusion of this section.

Accepted for publication in *Monthly Weather Review.* DOI 10.1175/MWR-D-22-0108.1.

2) RESULTS

We summarize the results for the "normal" setting in Fig. 3, and the mean values of κ in the adaptive experiments with this setting are shown in Fig. 8a. In Linear Case (Fig. 3a), when the number of members is small, the higher the ratio of the EnKF, or the closer the value of κ is to zero, the lower the RMSE. However, with 40 members, the experiment that performs partial local PF is optimal ($\kappa = 0.3$); after 40 members, the performance of the EnKF hardly improves as the number of members increases, and in the experiments with 300 members, pure EnKF shows the worst score. Here, since the sampling error decreases as the number of particles increases, it would seem that using more of the local PF update would give better results. However, for this particular model and a linear observation operator, this is not the case. Even with $N_e = 300$, the best performing experiments use a factorization that amounts to 70% of the EnKF increment being used. The experiment that determines the value of κ adaptively shows less optimal but suitable results for smaller ensemble sizes—with the added benefit of not needing to be tuned. As the number of members increases, however, the sample size for SWT increases, thus making the test more accurate. Increasing the ensemble size also increases the rejection rate of the null hypothesis, which is a desirable property. While the mean value of κ becomes larger as the ensemble size increases, the value converges slowly to 0.2 (Fig. 8a). The L96 priors remain close to Gaussian for most data assimilation cycles when using a sufficiently dense network of observations with linear measurement operators, thus leading us to conclude that SWT operates appropriately for this application.

Results from Nonlinear Case 1 are shown in Fig. 3b. With $N_e = 20 - 40$, the trend is the same as in the Linear Case: the experiments with a more significant percentage of the EnKF show better scores. However, this feature is maintained even with large ensemble sizes. This is because of the precision and frequency of the observations compared to ones in Linear Case, as described in Kurosawa and Poterjoy (2021). Since model variables are around the magnitude of $O(10)$, the nonlinear operator $H(\mathbf{x}) = \hat{\mathbf{x}} \circ \hat{\mathbf{x}}$ with $\sigma_y = 1.0$ provides very precise information to characterize the posterior estimation. This fact, combined with the frequency of measurements, makes Gaussian estimation more appropriate, as forecasts yield prior members that are generally close to the truth. Therefore, we can confirm that κ in the adaptive experiment uses a larger percentage of the EnKF than in Linear Case in Fig. 8a.

Accepted for publication in *Monthly Weather Review.* DOI 10.1175/MWR-D-22-0108.1.

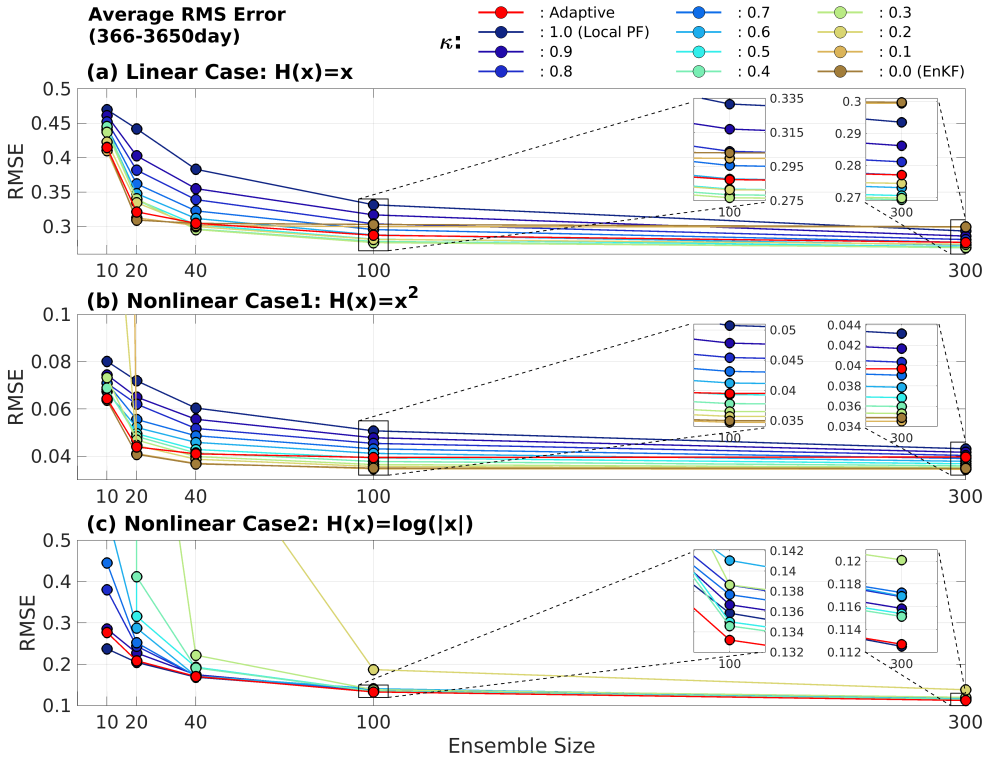

FIG. 3. Mean analysis RMSEs of the 12 experiments with different settings of κ as a function of ensemble size. Results are shown for (a) Linear Case, (b) Nonlinear Case1, and (C) Nonlinear Case2. Values are from the average of the last 9 years with 100 parallel trials.

Figure 3c shows the mean RMSEs from experiments that use measurements simulated with Nonlinear Case 2. In experiments using this observation network, a situation occurs in which the nonlinearity in the application becomes much larger than the sampling error in the prior and posterior distributions estimated by the ensemble. Owing to the strong nonlinearity of the observations, experiments using mainly the Gaussian-based method struggle to provide an accurate RMSE. In particular, the pure EnKF diverges, even with $N_e = 300$. The mean value of κ in the adaptive experiment shows that most of the update is used for the local PF (Fig. 8a). This result occurs as the strongly nonlinear observation operator tends to induce skewness in prior distributions, and SWT frequently rejects the null hypothesis.

Based on the above results from the "normal" setting, results from "mixed" observation networks yield intuitive results (Fig. 4). For example, in the case where the observation operators $H(\mathbf{x}) = \hat{\mathbf{x}}$ and $H(\mathbf{x}) = \hat{\mathbf{x}} \circ \hat{\mathbf{x}}$ are combined, the experiments with larger values of κ tend to produce worse

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-22-0108.1.

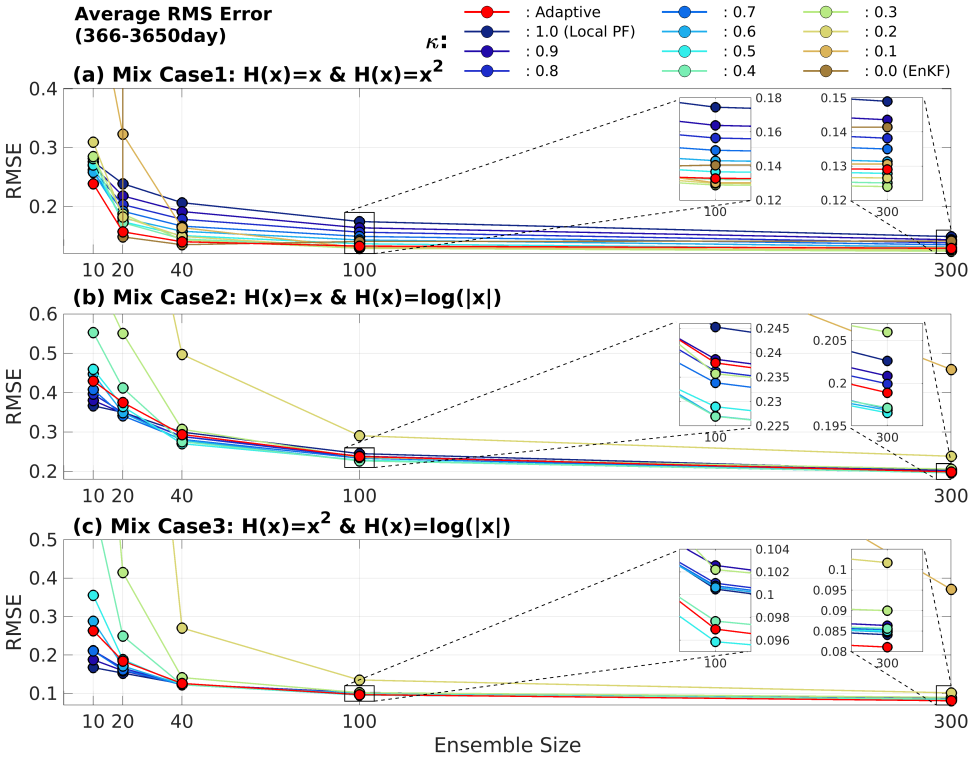

FIG. 4. As in Fig. 3, but for (a) Mix Case1, (b) Mix Case2, and (c) Mix Case3.

scores (Fig. 4a). On the other hand, experiments using strongly nonlinear operators experienced the best performance with small values of κ, and became unstable when the EnKF contribution was too large. As such, we consistently find that the experiments with a value of κ close to 0.5 are very stable. The partial update by the local PF adjusts particles to a Gaussian-like distribution, providing an optimal prior distribution for the EnKF update. The adaptive experiment also shows a satisfactory performance in the case of any combination of observation operators. We can see that using SWT is able to estimate the optimal κ according to each observation operator.

We summarize the results for the "data void" setting in Fig. 5. Mean RMSEs are uniformly higher than the "normal" setting with any observation operators despite using the smaller localization scale. Notably, in Nonlinear Case 2, several fixed experiments diverged, but the experiments with the appropriate blend of the EnKF and the local PF (κ=0.3-1.0) are stable (Fig. 5c). The mean value of κ used in the adaptive experiment is close to 1 (Fig. 8b). We note that this experiment shows slight advantages over experiments that keep κ fixed near 1, thus underscoring the importance of allowing κ to change over space and time.

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-22-0108.1.

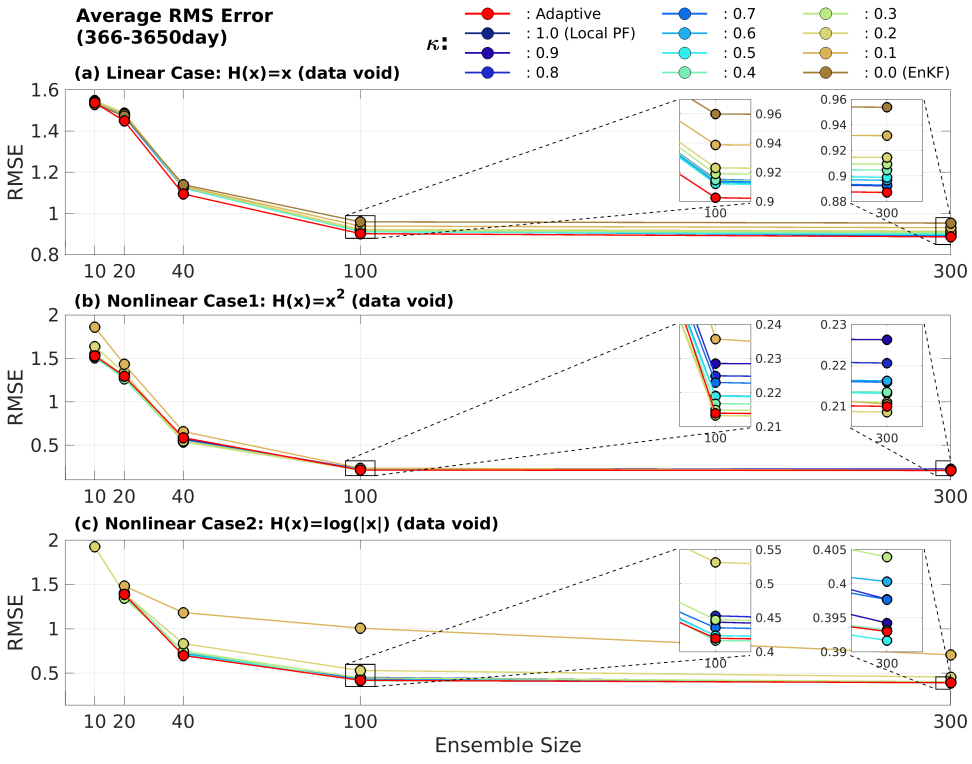

FIG. 5. As in Fig. 3, but using the data void observation network.

Lastly, results obtained from simulated "model error" experiments show elevated errors for all experiments, regardless of observation operators in Fig. 6. The presence of model errors means that the prior variance can be quite large, which leads to more frequent non-Gaussian prior distributions for L96. The hybrid strategies with a value of κ close to 0.5 show clear advantages in this regime. The parametric (Gaussian) assumption that follows the PF steps in hybrid configurations allows the filter to more easily adjust solutions for observations that lie outside the span of the ensemble. Hence, it shifts particles closer to observations in a manner that is not permitted by the PF—for variables that are detected to have Gaussian errors.

To investigate the behavior of SWT specification of κ for these simulations, we examine a sample time series of prior ensemble variance and estimated κ for the experiment using a linear measurement operator (Fig. 7). The plotted values come from the first variable of the L96 model in the first trial of the experiments with ensemble size $N_e = 300$. Because of an imperfect model, the prior variance fluctuates significantly over the entire period. When the prior distribution has a larger variance, the nonlinear model dynamics can more readily produce non-Gaussian priors, which SWT

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-22-0108.1.

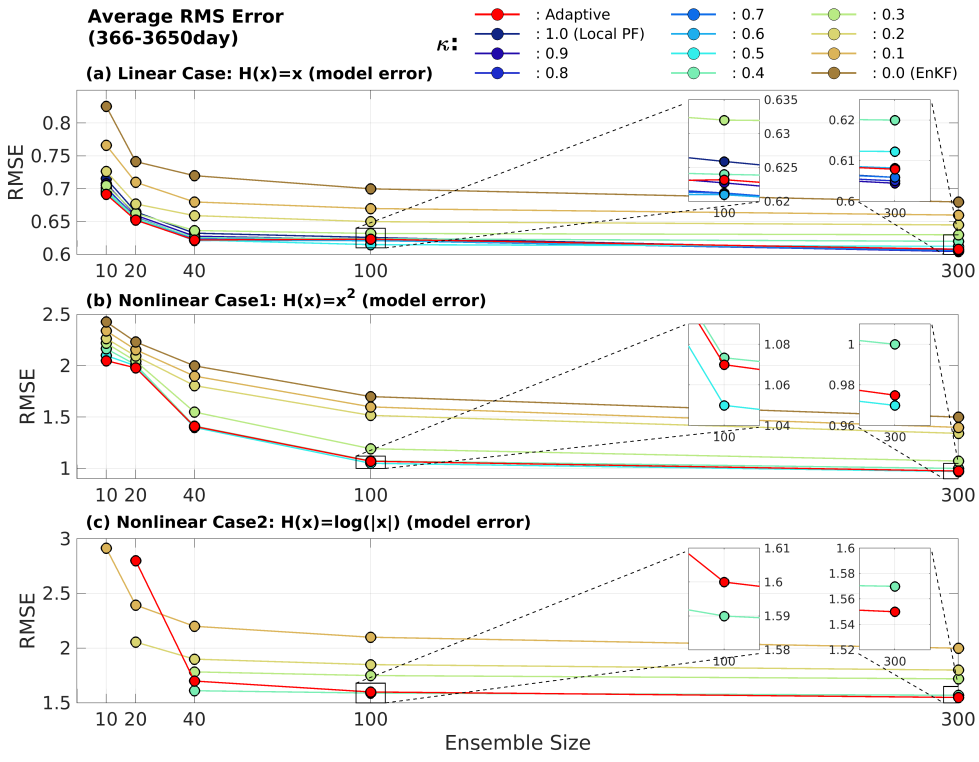

FIG. 6. As in Fig. 3, but using an imperfect L96 model for forecast steps.

successfully detects. For this example, the only factor that can contribute to non-Gaussian priors is the nonlinear model itself as the measurement operator is linear. Hence, fluctuations of the ensemble variance and κ are highly correlated. In Fig. 7, the time- and space-average value of κ over the period is 0.2655, which is very close to the value 0.2766 in Fig. 8b (note that the value of κ in Fig. 7 is from the first variable in the L96 model, while in Fig. 8 is from the average of all variables in the model). As in the "data void" simulations, the experiments with adaptively-estimated κ again show improvements over experiments with values of κ that are configured to use close to the average mean estimated κ, but fixed over space and time (Fig. 6a). In general, we find that choosing κ adaptively is beneficial in "model error" experiments, owing to its ability to maintain filter stability without rigorous tuning. The sporadic non-Gaussian priors produced by L96 in "model error" experiments introduce a major challenge that mimics the expected behavior of real weather systems.

Based on the above results, the statistical hypothesis testing approach yields adequate hybrid factor estimates in all situations we examined for this study. Moreover, the approach has significant

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-22-0108.1.

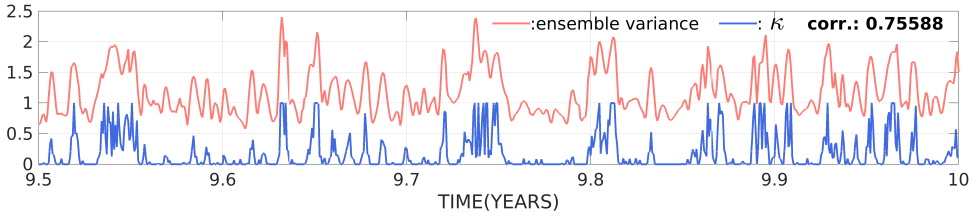

FIG. 7. Time series of ensemble spread (red) and estimated κ (blue) in adaptive hybrid experiments for Linear Case with "model error" setting. Values are from the first variable of the L96 model in the first trial of the experiments with ensemble size $N_e = 300$. The correlation coefficient between the pair of time series is represented in the top of right.

value for more realistic applications, such as non-homogeneous observation networks and unknown model process error. We expect similar benefits for geophysical problems that are characterized by a variety of dynamic instabilities as well. Furthermore, the proposed adaptive hybrid method avoids the need to tune heuristic parameters, such as the hybrid factor, which we find to be sensitive to observation operators, observation density, and model process uncertainty.

c. Idealized vortex model

In contrast to the low-dimensional applications used in the previous subsections, realistic atmospheric forecast models have several variables at each grid point, such as air temperature, winds, pressure, and specific humidity. These variables also exhibit large spatial error dependence with one another, which is not accounted for in adaptive choices for κ. As such, an observation of one

Accepted for publication in *Monthly Weather Review.* DOI 10.1175/MWR-D-22-0108.1.

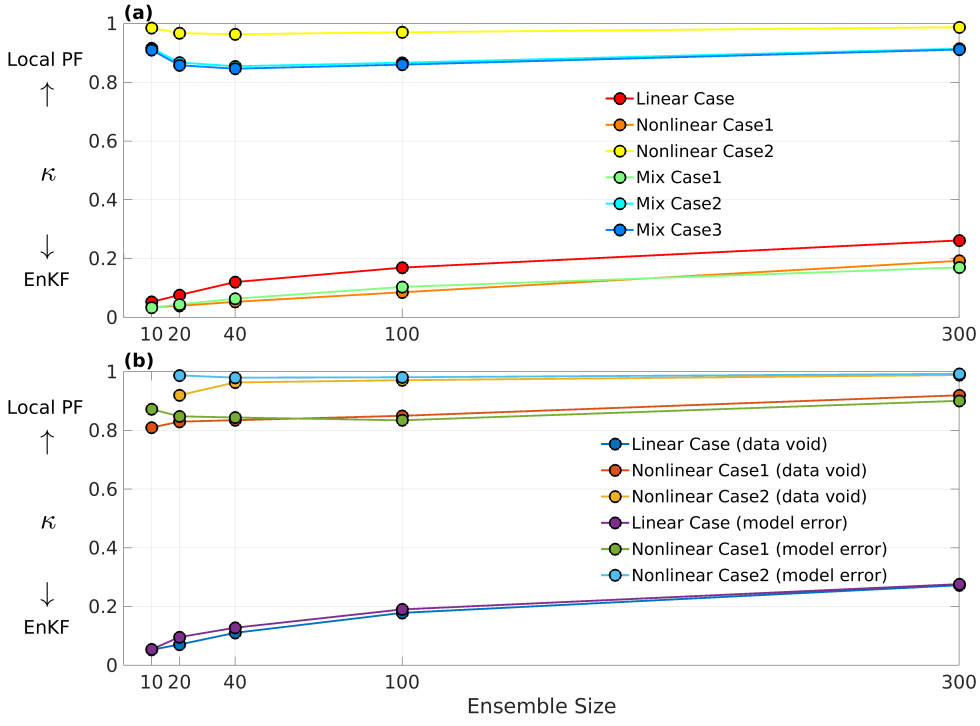

FIG. 8. Mean estimated κ of the adaptive hybrid experiments as a function of ensemble size. Results are shown for (a) "normal" and "mix" settings in Fig. 3 and Fig. 4, and (b) "data void" and "model error" settings in Fig. 5 and Fig. 6. Values are from the average of the last 9 years with 100 parallel trials.

variable must be used to update all co-located and nearby variables; for EnKFs, this step considers prior error covariance across each variable. Furthermore, extending κ to be the same dimension as the full state vector, rather than the grid dimension, would bring additional algorithmic complexity to the proposed hybrid filter. To address the problem of co-located variables for estimating κ, a natural choice is to perform the hypothesis test using all variables that are expected to be correlated with that variable at a grid point, i.e., by performing a test for multivariate normality. The numerical experiments performed in this section serve the purpose of illustrating the advantages of optimally adjusted κ estimated via SWT extended to the test for multivariate normality proposed by Royston (1983). Considering marginal PDFs in the test for multivariate normality is expected to provide a reasonable κ for data assimilation updates that account for correlations between co-located variables, which is an important practical feature of the proposed hybrid method in real weather applications, as the transition between the local PF and EnKF updates are decided across

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-22-0108.1.

grid points rather than state variables. The method, however, still neglects dependence for variables at different grid points, which is one theoretical shortcoming.

The current section provides an illustrative comparison of the EnKF, the local PF, and hybrid updates using a low dimensional application that mimics a common challenge for filtering geophysical flow, namely the problem of assimilating measurements for mesoscale weather features that are not well constrained by measurements.

1) EXPERIMENTAL DESIGNS

Adopting the same application introduced in Poterjoy (2022a), we will reproduce the data assimilation challenge posed by alignment errors associated with mesoscale weather features by modeling a vortex wind field with a Rankine vortex profile (Acheson 1990). The model produces a case of a single vortex in zero-mean flow, but position uncertainty. The Rankine vortex consists of a wind field exhibiting uniform vorticity in the vortex, and an outer region of zero vorticity. For cylindrical coordinates with the origin chosen to be the vortex center, and assuming that all non-zero vorticity is uniformly distributed within a circle of radius R_{max}, the tangential winds u_θ are a function of radius r:

$$u_\theta = \begin{cases} U_\theta \frac{r}{R_{max}}, r < R_{max} \\ U_\theta \frac{R_{max}}{r}, r \geq R_{max} \end{cases}, \tag{14}$$

where U_θ is maximum wind speed. Both the radial wind component (u_r) and vertical wind component (u_z) are assumed to be zero. For this demonstration, we transform winds into Cartesian coordinates so that the model state vector is comprised of zonal (u) and meridional (v) wind components; i.e., $\mathbf{x} = [\mathbf{u}, \mathbf{v}]^\intercal$.

The current study generates vortices on a two-dimensional Cartesian region consisting of 91 × 91 equally spaced grid points. To generate a prior sample, we first designate a control state with the center of the vortex located at ($i_c^{CTRL} = 46, j_c^{CTRL} = 46$), and $U_\theta^{CTRL} = 30 ms^{-1}$ and $R_{max}^{CTRL} = 12$. Then, the position and wind parameters of each vortex are randomly drawn independently from a Gaussian distribution, and added to the control vortex parameters. That is, the center of each vortex (i_c^n, j_c^n) is sampled from $N(i_c^{CTRL}, \sigma_p^2)$ and $N(j_c^{CTRL}, \sigma_p^2)$, respectively, for $n = 1, ..., N_e$, where σ_p is a prescribed position error standard deviation that changes for each prior. U_θ^n and R_{max}^n of each vortex is drawn from $N(U_\theta^{CTRL}, 1)$ and $N(R_{max}^{CTRL}, 1)$, respectively.

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-22-0108.1.

Observations are generated by uniformly selecting points from within the scan area of a hypothetical Doppler radar with a radius of 30 grids, placed at coordinates ($i_{radar} = 25, j_{radar} = 25$) in the lower-left corner of the domain. We produce each observation by projecting the truth state wind in the direction of the hypothetical radar beam pointing outward from the radar. In this experiment, the errors added to each observation are drawn from $N(0, \sigma_o^2)$ for $\sigma_o = 3$, and we set the number of observations N_y to be 100. Figure 9a shows the value of u_θ for the cross section through the center of the control state. Reproduced from Poterjoy (2022a), Fig. 9b shows a single $15ms^{-1}$ wind speed contour for the Rankine vortex on the 2-D domain. The scan region by the virtual radar is indicated by the curved segment in the lower left part of the domain, which covers only one quadrant of the vortex, and the green and red dots indicate the location and magnitude of the measurements (Fig. 9b).

As in the experiment with the L96 in Section 4b, out of an abundance of caution, we perform 3000 parallel trials with unique sets of priors, true solutions, and observations in order to capture the range of plausible outcomes for this application. To perform each trial, for the truth state, we generate the center (i_c^t, j_c^t), maximum wind speed U_θ^t, and radius R^t from a Gaussian distribution as well as other prior ensemble members. This process allows us to create a state in which the true value is indistinguishable from any prior member, i.e., the true value is also a sample from the prior distribution with equal probability, which is a condition assumed when performing data assimilation for real atmospheric models. The truth state, which varies by each trial, is used to generate the observations and provides a reference for evaluating data assimilation experiments performed for each trial. We repeat these trials for several choices of position error standard deviation $\sigma_p = \{0.0, 4.0, 8.0, 12.0\}$ and ensemble size $N_e = \{40, 100, 300\}$. For reference, Fig. 10 shows the variability of the initial ensemble members according to each σ_p with $N_e = 40$.

We performed a total of 12 experiments, one in which the value of κ is estimated adaptively, and the others in which κ is spatially constant between 0 and 1, using increments of 0.1. In the adaptive hybrid experiment, we use SWT extended to the test for multivariate normality; that is, the Gaussianity in the prior samples is detected using two variables, u and v. In this demonstration, we assimilate the observations in each of the κ settings using each of the choices of prior and then calculate the RMSEs of the posterior mean relative to the true wind velocity. All experiments use

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-22-0108.1.

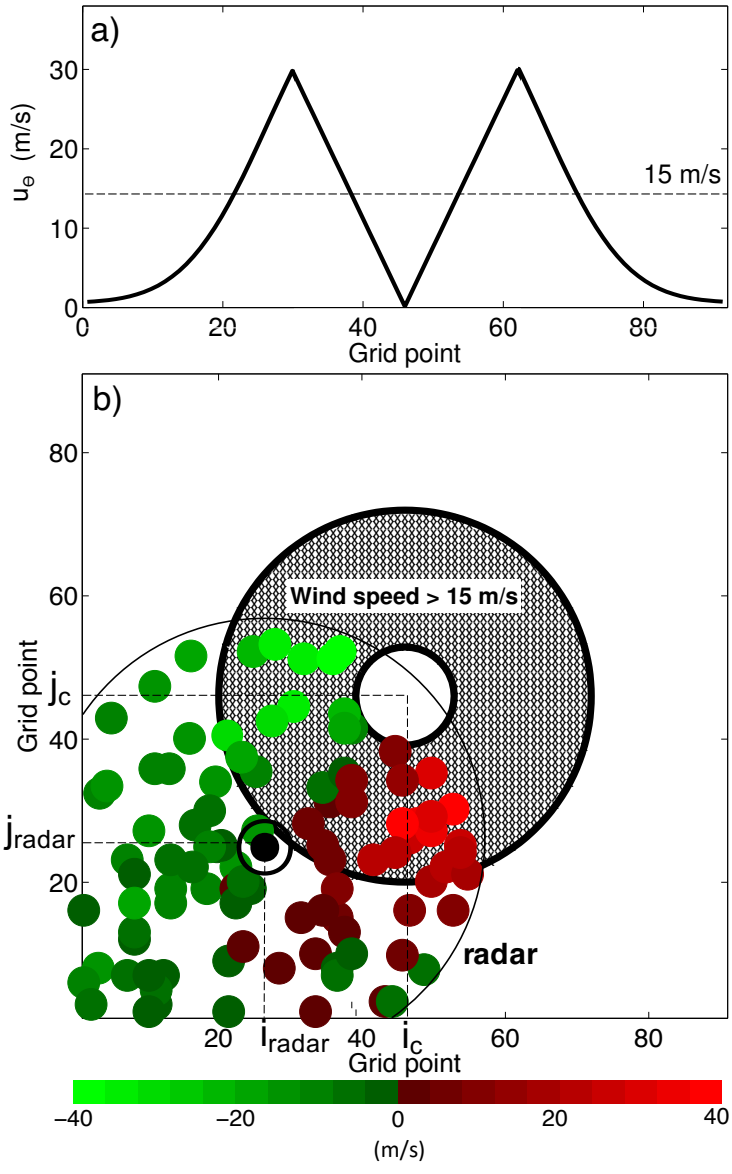

FIG. 9. Top panel (a) is tangential wind speed as a function of grid points, calculated using the Rankine vortex model with point 46 as center location. The bottom panel (b) shows 15 m/s wind speed contours for the vortex placed on a 2-D grid; values greater than 15 m/s are indicated by hatched region. The green and red markers indicate the location and magnitude of radial wind observations created for a synthetic radar located at coordinate (i_{radar}, j_{radar}). This figure is a reproduction of Fig.8 in Poterjoy (2022a).

the localization function f:

$$f = \exp\left\{-\frac{1}{2}\left[\frac{d(i,j)}{\sigma_{loc}}\right]^2\right\}, \tag{15}$$

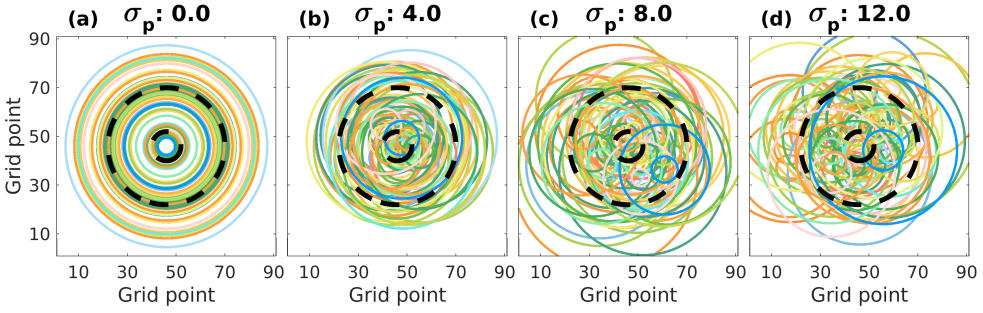

FIG. 10. Variability of the initial ensemble members for (a) $\sigma_p = 0.0$, (b) $\sigma_p = 4.0$, (c) $\sigma_p = 8.0$, and (d) $\sigma_p = 12.0$ with ensemble size $N_e = 40$. Each colored line shows the 15m/s wind speed contours. The black dashed lines show the 15m/s wind speed contours of the control state.

where $d(i, j)$ is the physical distance between grid points i and j, and σ_{loc} is the localization parameter to scale the width of localization, which is set to 2000 in the current study.

2) RESULTS

This subsection discusses results obtained by performing single-cycle data assimilation using the described sets of observations and prior members. Figure 11 shows the posterior RMSEs for four experiments as examples: EnKF ($\kappa = 0.0$), the local PF ($\kappa = 1.0$), PF-EnKF with $\kappa = 0.5$, and PF-EnKF with adaptive κ estimation. For experiments that use $\sigma_p = 0.0$ for the prior, each data assimilation method shows low RMSEs that are visibly similar (first row of Fig. 11). This finding is expected because the small $\sigma_p = 0.0$ leads to the Gaussian assumption being valid (Poterjoy 2022a). However, as the value of σ_p increases, each experiment yields vast differences in the upper right portion of the domain where each filter must infer wind estimates from distant

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-22-0108.1.

observations. This application is especially problematic for the EnKF, as linear updates do not properly capture nonlinear dependence in winds across the vortex (Poterjoy 2022a). While the local PF and PF-EnKF produce smaller mean RMSEs than the EnKF, we note that these errors also continue to decrease as the number of members increases, because of the decrease in sampling error. Furthermore, compared to the EnKF and local PF experiments, both hybrid experiments show more accurate results across all domains, which demonstrates that the hybrid PF-EnKF method is effective at shifting particles into an approximate Gaussian before applying the EnKF step—even for the highly non-Gaussian vortex application discussed in Poterjoy (2022a).

Comparing experiments with fixed and adaptive κ, Fig. 12 shows the grid points over the domain where a specified value of κ produced the smallest RMSEs for each choice of prior. For all settings, the adaptive estimate yields the smallest errors outside of the vortex, thus reflecting diversity in optimal κ in this region. When $\sigma_p = 0$, experiments with a value of κ fixed near 0.5 show the best results near the vortex center (Fig. 12a-c). We suspect this result occurs because the location of the prior and true vortex centers are the same for all prior distributions when $\sigma_p = 0$, but since R is drawn from $N(R^{CTRL}, 1)$, the winds exhibit bimodal behavior, which is controlled by parameter R in (14); recall, this parameter divides the domain into regions of zero and non-zero—but constant—vorticity. Since the region near the vortex center is characterized by the presence of both zeros and nonzeros, it is conceivable that the case of $\kappa = 0.5$, where both PF and EnKF can be used in a balanced manner, happens to be the most optimal. Therefore, a fixed value of κ can be identified via rigorous tuning, rather than resorting to hypothesis testing. When $\sigma_p > 0$ and a sufficiently large ensemble size is used (e.g., $N_e = 300$), SWT correctly identifies values for κ that outperform fixed values for κ over most of the domain (Figs. 12f,i,l). Prior vortices are no longer co-located as σ_p increases, so the region where fixed values for κ are optimal gradually extends outward from the center.

We also examine the mean value of κ (averaged over trials) in the experiments where κ is adaptively adjusted (Fig. 13). First, for the case of $\sigma_p = 0.0$, the area close to the center of the control vortex, where experiments with a value of κ fixed near 0.5 corresponds to the location where estimated κ is about 0.2-0.5 in the first row of Fig. 13. This indicates that the area has a low percentage of the local PF updates compared to the fixed experiment. For the cases of $N_e = 40$ and 100, the areas where the experiment using the adaptively adjusted κ is inferior in Fig.

Accepted for publication in *Monthly Weather Review.* DOI 10.1175/MWR-D-22-0108.1.

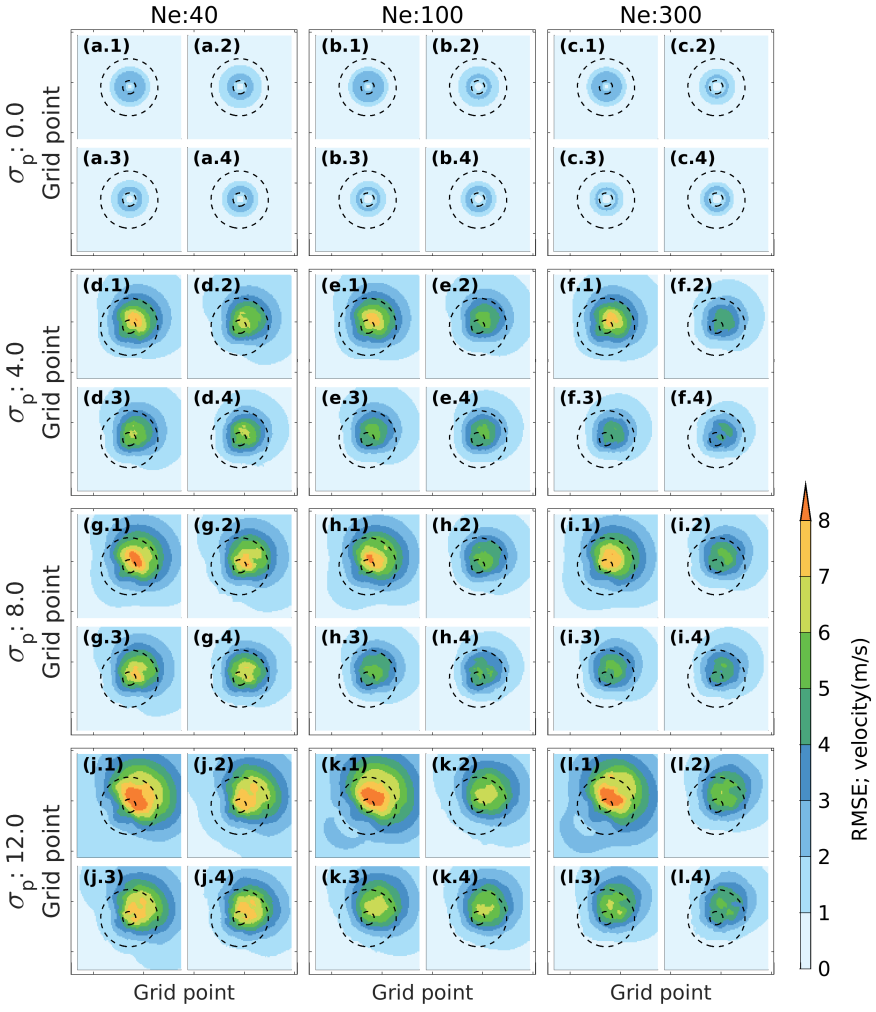

FIG. 11. Analysis RMSEs of velocity for (1) EnKF ($\kappa = 0.0$), (2) PF ($\kappa = 1.0$), (3) the experiment with $\kappa = 0.5$, and (4) with adaptive estimated κ. The position error standard deviation σ_p is 0.0, 4.0, 8.0, 12.0 for first row (a-c), second row (d-f), third row (g-i), and forth row (j-l), respectively. The ensemble size N_e is 40, 100, and 300 for first column (a,d,g,j), second column (b,e,h,k), and third column (c,f,i,l), respectively. Values are from the average of 3000 parallel trials. The black dashed lines show the 15m/s wind speed contours of the control state.

12 are generally estimated to have a value of κ less than 0.5 (the first two columns of Fig. 13). However, in the case with $N_e = 300$, the values of κ in those locations are generally more than 0.5, and the difference from the fixed experiment is not significant (third column of Fig. 13). This may be because the larger sample size used in SWT leads to more frequent rejection of the null

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-22-0108.1.

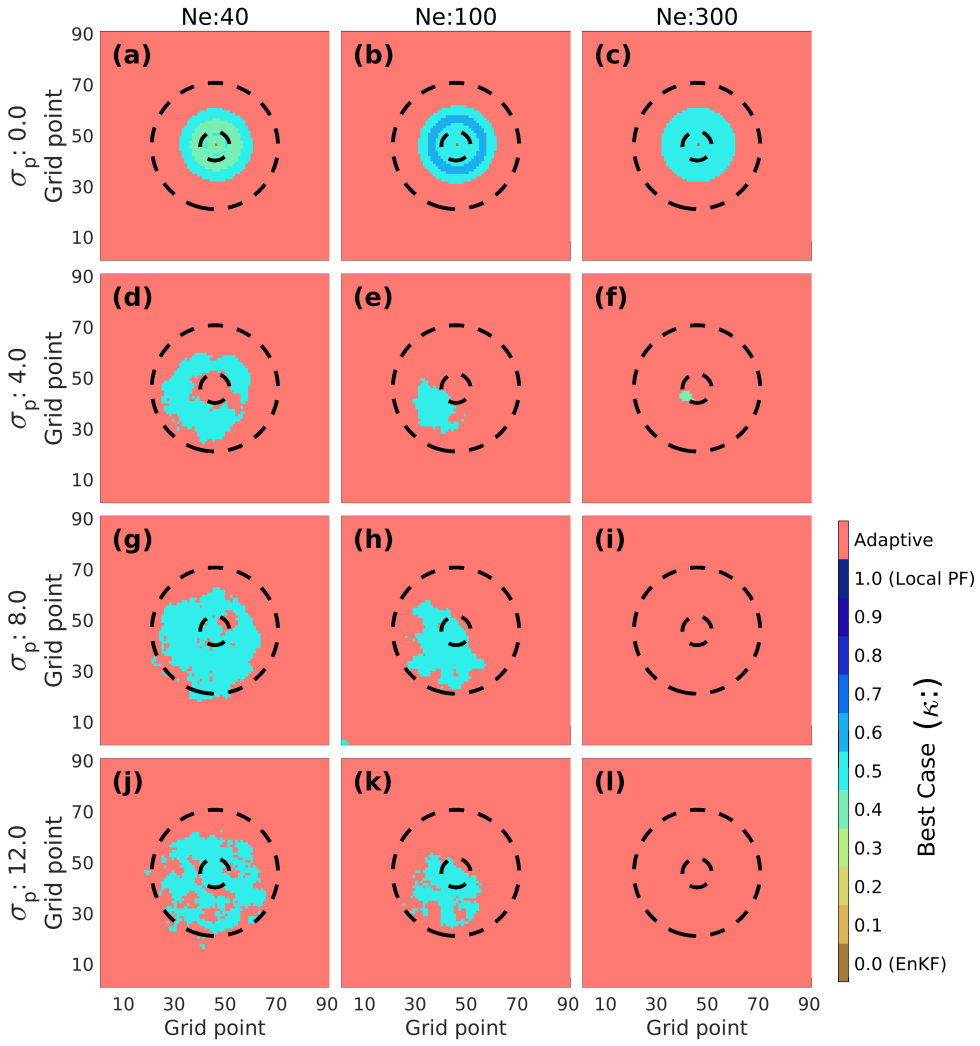

FIG. 12. The experiment with the lowest RMSEs out of the 12 different κ cases over 3000 parallel trials. Color represents the best κ case, which produced the smallest RMSE, comparing fixed and adaptive κ experiments. The black dashed lines show the 15m/s wind speed contours of the control state.

hypothesis, resulting in more iterations of the local PF. Furthermore, the experiment with estimated κ is generally more stable in the other areas, far away from the center, especially in the upper right domains, where there are no observations.

Accepted for publication in *Monthly Weather Review.* DOI 10.1175/MWR-D-22-0108.1.

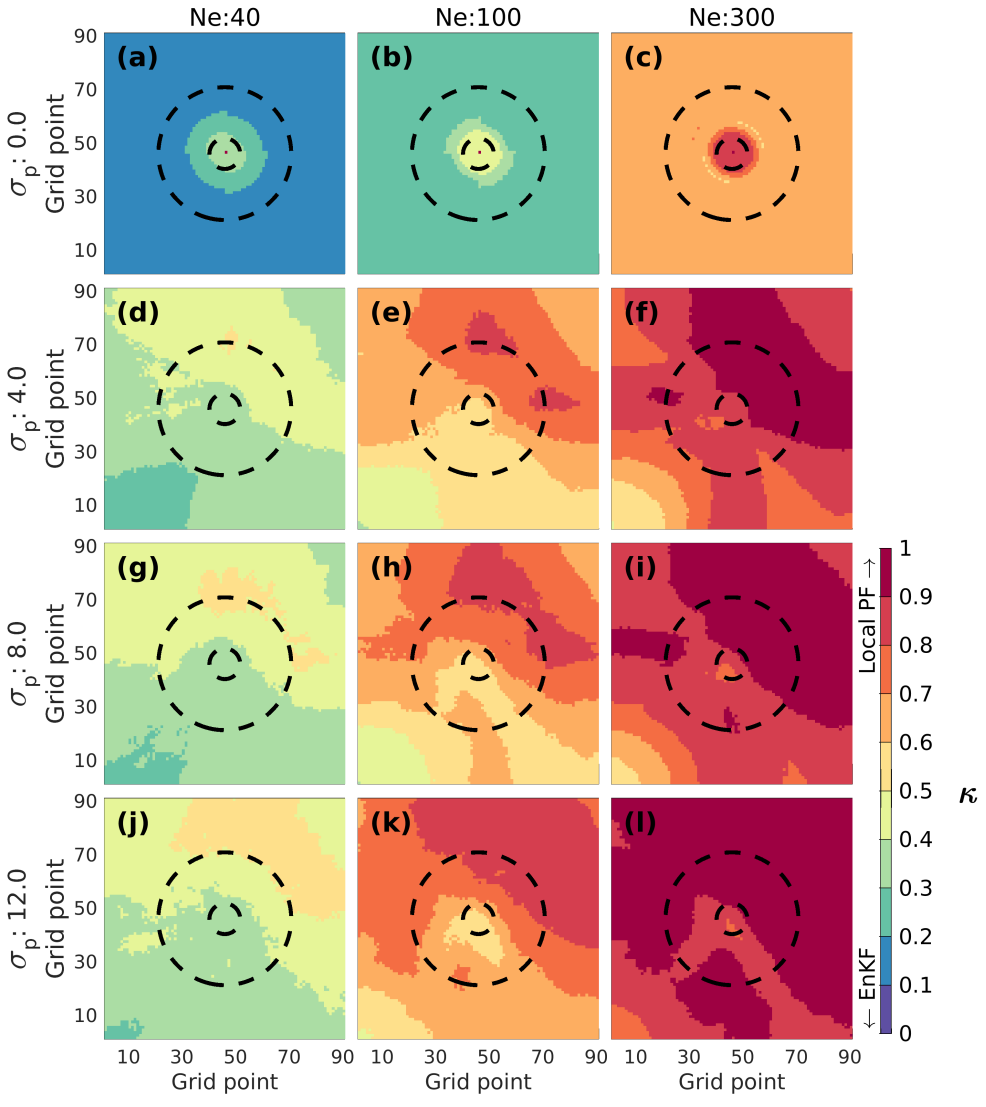

FIG. 13. Mean value of κ in the experiments with adaptive estimated κ. Values are from the average of 3000 parallel trials. The black dashed lines show the 15m/s wind speed contours of the control state.

As in the previous subsections, the kinematic vortex experiments illustrate the advantage of specifying κ adaptively versus keeping this parameter fixed. In this demonstration, however, κ is estimated using u and v with SWT extended to handle multivariate normality. Hence, results from this section differ in that we successfully analyzed samples from multivariate probability distri-

Accepted for publication in *Monthly Weather Review.* DOI 10.1175/MWR-D-22-0108.1.

butions. The multivariate approach is more practical for data assimilation with real atmospheric models, when we need to consider correlation among co-located variables. Furthermore, as noted by Poterjoy (2022a), the non-Gaussian data assimilation problem constructed in this section has univariate marginal distributions that are close to Gaussian, but multivariate marginals for variables across grid points that are far from Gaussian. In terms of computing burden, high-resolution models, such as those used for weather forecasting, are constrained by ensemble size. As a result, detecting such characteristics with a limited number of ensembles to be utilized for operational models is extremely challenging. Nevertheless, we find the proposed SWT approach to be sufficient for identifying deviations from a multivariate normality for co-located winds, which shows added value over a rigorously tuned hybrid methodology that uses fixed specifications for κ.

5. Discussion and conclusions

The current study introduces a novel approach to forming an adaptive hybrid data assimilation method that mixes the theoretical strengths and flexibility of particle filters with Gaussian-based ensemble Kalman filters (EnKFs), which are more resilient to bias in sample-estimated prior uncertainty. For this purpose, we use a recently proposed PF by Poterjoy (2022b), which introduces a regularization and tempering methodology to improve filter performance when sampling error is large. The tempering step consists of a factorization of the particle weights, which provides a natural framework for combining local PFs with alternative filters to mitigate the effects of sampling error. In addition to identifying portions of the state space where a PF may provide more accurate marginal posterior estimates than an EnKF, the adaptive strategy can switch between filters partway through data assimilation steps. The latter property is beneficial when Gaussian assumptions are appropriate for posteriors but not for priors, which is common when likelihoods are Gaussian. In this case, partial updates performed by the PF can adjust the distribution of particles to more closely fit a Gaussian, which allows for a more effective use of EnKFs. To determine the timing of the transition between these filter updates, we use the Shapiro-Wilk test (SWT), which has excellent power among omnibus tests to detect deviations from normality. The use of SWT allows for accurate detection of Gaussianity even when the ensemble size is small. SWT also requires minimal computing time, thus permitting its use between PF iterations, which can be carried out until prior sample distributions for marginals at each grid point are detected to

Accepted for publication in *Monthly Weather Review.* DOI 10.1175/MWR-D-22-0108.1.

be Gaussian. Increasing the ensemble size also increases the rejection rate of the null hypothesis and leads to a smaller portion of updates being made by an EnKF, which is a desirable property.

To examine the performance of the adaptive hybrid, this study constructs numerous data assimilation experiments using a low-dimensional dynamical model, which is characterized by 40 equally-spaced variables on a periodic domain. In general, the statistical hypothesis testing approach yields adequate estimates of the hybrid factor in all situations considered in this study. Given a homogeneous network of equally-spaced observations, the adaptive formulations of the hybrid filter are as accurate as the rigorously tuned hybrid parameters. The adaptive approach also demonstrates clear advantages in experiments containing heterogeneous observation networks and unknown model process errors—in which case, the optimal choice of adaptive parameter varies temporally or across variables.

The study also examines practical challenges for adopting the new method for real Earth system models, which are characterized by multiple variables at common grid points and large error correlations through space; e.g., modern weather prediction models. The computational expense of such models limits the amount of tuning that can be performed for heuristic parameters used during data assimilation, which can be sensitive to observation operators, observation frequency, and model process uncertainty. Therefore, this study adopts an idealized kinematic vortex model to study the behavior of the adaptive hybrid. This model permits large error dependence across variables displaced over a two-dimensional domain, and contains two variables (zonal and meridional wind) at each grid point thus requiring a multivariate SWT to adaptively choose how to partition PF and EnKF updates. For this application, the hybrid factor is estimated using SWT extended to detect multivariate normality for ensembles of u and v at each grid point. This approach allows the use of the appropriate factor to account for multivariate marginal distributions for updating the state variables, alongside observation-space priors. Specifying the hybrid factor for co-located variables also simplifies the algorithmic formulation of the adaptive methodology, as it only requires the factor to be specified for all grid points and observation-space priors used during data assimilation. The experiments reveal spatial patterns of adaptively-chosen hybrid factors that result in large PF updates in portions of the state space where Gaussian assumptions are known to be incorrect, and are close to the values identified at each grid point from rigorously-tuned experiments aimed

Accepted for publication in *Monthly Weather Review.* DOI 10.1175/MWR-D-22-0108.1.

at reducing posterior mean RMSEs. These results encourage further testing for real geophysical problems that are characterized by a variety of dynamic instabilities.

In summary, the proposed adaptive hybrid method performs well in idealized simulations that mimic data assimilation problems encountered for real geophysical modeling systems. Because the new strategy relies on statistical hypothesis testing, it becomes more stable when the ensemble size increases. The proposed method obviates the need for tuning a hybrid parameter that influences when an EnKF is preferred over PF, which can depend on a number of factors including the underlying model dynamics and observation network. This property of the method has theoretical benefits for real Earth system models where rigorous tuning of data assimilation parameters is not always feasible, and the shape of error distributions is flow-dependent. Lastly, this study demonstrates how SWT can be extended to consider error dependence for co-located variables. Further research will explore the use of multivariate error dependence for variables across grid points, which may be needed for prior distributions that are characterized by strong nonlinear dependence for variables displaced geographically.

Accepted for publication in *Monthly Weather Review.* DOI 10.1175/MWR-D-22-0108.1.

Acknowledgments. Funding for this work was provided by NOAA grant #NA20OAR4600281 and NSF/CAREER Award #AGS1848363.

Accepted for publication in *Monthly Weather Review.* DOI 10.1175/MWR-D-22-0108.1.

Data availability statement. This paper uses numerous low-dimensional model simulations that can be easily replicated using software written for this study. The model code, compilation script, and the namelist settings are available at `https://github.com/Kenta9638/MWR_2022`.

Accepted for publication in *Monthly Weather Review.* DOI 10.1175/MWR-D-22-0108.1.

References

Acheson, D. J., 1990: Elementary fluid dynamics. *Oxford University Press*.

Althouse, L., W. Ware, and J. M. Ferron, 1998: Detecting departures from normality: A monte carlo simulation of a new omnibus test based on moments. *Paper presented at the Annual Meeting of the American Educational Research Association*, 33 pp, URL https://eric.ed.gov/?id=ED422385.

Anderson, T. W., and D. A. Darling, 1954: A test of goodness of fit. *Journal of the American Statistical Association*, **49 (268)**, 765–769.

Arshad, M., M. Rasool, and M. Ahmad, 2003: Anderson darling and modified anderson darling tests for generalized pareto distribution. *Journal of Applied Sciences*, **3(2)**.

Bocquet, M., C. A. Pires, and L. Wu, 2010: Beyond gaussian statistical modeling in geophysical data assimilation. *Monthly Weather Review*, **138 (8)**, 2997 – 3023, doi:10.1175/2010MWR3164.1, URL https://journals.ametsoc.org/view/journals/mwre/138/8/2010mwr3164.1.xml.

Chustagulprom, N., S. Reich, and M. Reinhardt, 2016: A hybrid ensemble transform particle filter for nonlinear and spatially extended dynamical systems. *SIAM/ASA J. Uncertainty Quantification*, **4(1)**, 592–608.

Cramér, H., 1928: On the composition of elementary errors. *Scandinavian Actuarial Journal*, **1928 (1)**, 13–74.

Dufour, J., A. Farhat, L. Gardiol, and L. Khalaf, 1998: Simulation-based finite sample normality tests in linear regressions. *The Econometrics Journal*, **1 (1)**, C154–C173.

Emerick, A. A., and A. C. Reynolds, 2012: History matching time-lapse seismic data using the ensemble Kalman filter with multiple data assimilations. *Comput. Geosci.*, **16(3)**, 639–659.

Errico, R., P. Bauer, and J.-F. Mahfouf, 2007: Issues regarding the assimilation of cloud and precipitation data. *J. Atmos. Sci.*, **64**, 3785–3798.

Evensen, G., 1994: Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *J. Geophys. Res.*, **99C**, 10 143–10 162.

Evensen, G., and P. J. van Leeuwen, 2000: An ensemble Kalman smoother for non-linear dynamics. *Mon. Wea. Rev.*, **128**, 1852–1867.

Accepted for publication in *Monthly Weather Review.* DOI 10.1175/MWR-D-22-0108.1.

Fabry, F., and J. Sun, 2010: For how long should what data be assimilated for the mesoscale forecasting of convection and why? Part I: On the propagation of initial condition errors and their implications for data assimilation. *Mon. Wea. Rev.*, **138**, 242–255.

Farrell, P. J., M. Salibian-Barrera, and K. Naczk, 2007: On tests for multivariate normality and associated simulation studies. *Journal of Statistical Computation and Simulation*, **77 (12)**, 1065–1080, doi:10.1080/10629360600878449.

Frei, M., and H. R. Kunsch, 2013: Bridging the ensemble Kalman and particle filters. *Biometrika*, 1–20.

Gaspari, G., and S. E. Cohn, 1999: Construction of correlation functions in two and three dimensions. *Quart. J. Roy. Meteor.*, **125**, 723–757.

Grooms, I. G., and G. Robinson, 2021: A hybrid particle-ensemble kalman filter for problems with medium nonlinearity. *PLoS ONE*, **16**.

Honda, T., and Coauthors, 2018: Assimilating all-sky himawari-8 satellite infrared radiances: A case of typhoon soudelor (2015). *Mon. Wea. Rev.*, **146**, 213–229.

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

Kolmogorov, A. N., 1933: Sulla determinazione empirica di una legge di distribuzione. *Giornale dell' Instituto Italiano degli Attuari*, **4**, 83–91.

Kondo, K., and T. Miyoshi, 2019: Non-gaussian statistics in global atmospheric dynamics: a study with a 10 240-member ensemble kalman filter using an intermediate atmospheric general circulation model. *Nonlinear Processes in Geophysics*, **26 (3)**, 211–225, doi:10.5194/npg-26-211-2019, URL https://npg.copernicus.org/articles/26/211/2019/.

Kullback, S., and R. A. Leibler, 1951: On information and sufficiency. *The Annals of Mathematical Statistics*, **22 (1)**, 79–86, URL http://www.jstor.org/stable/2236703.

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 (7)**, 2369–2389.

Accepted for publication in *Monthly Weather Review.* DOI 10.1175/MWR-D-22-0108.1.

Li, R., N. Magbool Jan, B. Huang, and V. Prasad, 2019: Constrained multimodal ensemble kalman filter based on kullback–leibler (kl) divergence. *Journal of Process Control*, **79**, 16–28, doi:https://doi.org/10.1016/j.jprocont.2019.03.012, URL https://www.sciencedirect.com/science/article/pii/S0959152418301756.

Lilliefors, H. W., 1967: On the kolmogorov-smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*, **62 (318)**, 399–402.

Lorenz, E. N., 1996: Predictability: A problem partly solved. *Proc. Seminar on Predictability*, **1**, ECMWF, Reading, UK.

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.

Mendes, M., and A. Pala, 2003: Type i error rate and power of three normality tests. *Information Technology Journal*, **2**, doi:10.3923/itj.2003.135.139.

Metref, S., E. Cosme, C. Snyder, and P. Brasseur, 2014: A non-gaussian analysis scheme using rank histograms for ensemble data assimilation. *Nonlinear Processes in Geophysics*, **21 (4)**, 869–885, doi:10.5194/npg-21-869-2014, URL https://npg.copernicus.org/articles/21/869/2014/.

Minamide, M., and F. Zhang, 2017: Adaptive observation error inflation for assimilating all-sky satellite radiance. *Mon. Wea. Rev.*, **145 (3)**, 1063–1081.

Morzfeld, M., and D. Hodyss, 2019: Gaussian approximations in filters and smoothers for data assimilation. *Tellus A: Dynamic Meteorology and Oceanography*, **71:1**, 1–27.

Morzfeld, M., D. Hodyss, and J. Poterjoy, 2018: Variational particle smoothers and their localization. *Quarterly Journal of the Royal Meteorological Society*, **144**.

Nerger, L., 2022: Data assimilation for nonlinear systems with a hybrid nonlinear kalman ensemble transform filter. *Quart. J. Roy. Meteor.*, **148 (743)**, 620–640, doi:https://doi.org/10.1002/qj.4221, URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.4221.

Penny, S. G., and T. Miyoshi, 2016: A local particle filter for high dimensional geophysical systems. *Nonlinear Processes in Geophysics*, **23**, 391–405.

Pimentel, S., and Y. Qranfal, 2021: A data assimilation framework that uses the kullback-leibler divergence. *PLOS ONE*, **16**, 1–18, doi:10.1371/journal.pone.0256584, URL https://doi.org/10.1371/journal.pone.0256584.

Poterjoy, J., 2016: A localized particle filter for high-dimensional nonlinear systems. *Mon. Wea. Rev.*, **144**, 59–76.

Poterjoy, J., 2022a: Implications of multivariate non-gaussian data assimilation for multi-scale weather prediction. *Mon. Wea. Rev.*, **150**, 1475–1493.

Poterjoy, J., 2022b: Regularization and tempering for a moment-matching localized particle filter. *Quart. J. Roy. Meteor. Soc.*, Published online.

Poterjoy, J., and J. L. Anderson, 2016: Efficient assimilation of simulated observations in a high-dimensional geophysical system using a localized particle filter. *Mon. Wea. Rev.*, **144**, 2007–2020.

Poterjoy, J., R. A. Sobash, and J. L. Anderson, 2017: Convective-scale data assimilation for the weather research and forecasting model using the local particle filter. *Mon. Wea. Rev.*, **145**, 1897–1918.

Poterjoy, J., L. J. Wicker, and M. Buehner, 2019: Progress in the development of a localized particle filter for data assimilation in high-dimensional geophysical systems. *Mon. Wea. Rev.*, **147**, 1107–1126.

Potthast, R., A. Walter, and A. Rhodin, 2019: A localized adaptive particle filter within an operational NWP framework. *Mon. Wea. Rev.*, **147**, 345–362.

Privé, N. C., Y. Xie, J. Woollen, S. E. Koch, R. Atlas, and R. Hood, 2013: Evaluation of the Earth Systems Research Laboratory's global Observing System Simulation Experiment system. *Tellus*, **65A**, 19 011.

Razali, N., and Y. Wah, 2011: Power comparison of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson darling tests. *Journal of Statistical Modeling and Analytics*, **2**, 21–33.

Royston, J. P., 1982: An extension of shapiro and wilk's w test for normality to large samples. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **31 (2)**, 115–124.

Accepted for publication in *Monthly Weather Review.* DOI 10.1175/MWR-D-22-0108.1.

Royston, J. P., 1983: Some techniques for assessing multivarate normality based on the shapiro-wilk w. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **32 (2)**, 121–133.

Royston, J. P., 1992: Approximating the shapiro-wilk w-test for non-normality. *Statistics and Computing*, **2**, 117–119.

Royston, J. P., 1995: Remark as r94: A remark on algorithm as 181: The w-test for normality. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **44 (4)**, 547–551.

Ruiz, J., G.-Y. Lien, K. Kondo, S. Otsuka, and T. Miyoshi, 2021: Reduced non-gaussianity by 30 s rapid update in convective-scale numerical weather prediction. *Nonlinear Processes in Geophysics*, **28 (4)**, 615–626, doi:10.5194/npg-28-615-2021, URL https://npg.copernicus.org/articles/28/615/2021/.

Saculinggan, M., and E. A. Balase, 2013: Empirical power comparison of goodness of fit tests for normality in the presence of outliers. *Journal of Physics Conference Series*, **435**, 2041–2052, doi:10.1088/1742-6596/435/1/012041, URL https://doi.org/10.1088/1742-6596/435/1/012041.

Shapiro, S. S., and M. B. Wilk, 1965: An analysis of variance test for normality (complete samples). *Biometrika*, **52 (3/4)**, 591–611, URL http://www.jstor.org/stable/2333709.

Slivinski, L., E. Spiller, A. Apte, and B. Sandstede, 2015: A hybrid particle-ensemble Kalman filter for Lagrangian data assimilation. *Mon. Wea. Rev.*, **143**, 195–211.

Smirnov, N. V., 1939: Estimate of deviation between empirical distribution functions in two independent samples. *Bulletin Moscow University*, **2 (2)**, 3–16.

Srivastava, M., and T. Hui, 1987: On assessing multivariate normality based on shapiro–wilk w statistic. *Statistics and Probability Letters*, **5**, 15–18, doi:10.1016/0167-7152(87)90019-8.

Stengel, M., P. Unden, M. Lindskog, P. Dahlgren, N. Gustafsson, and R. Bennartz, 2009: Assimilation of SEVIRI infrared radiances with HIRLAM 4D-Var. *Quart. J. Roy. Meteor. Soc.*, **135**, 2100–2109.

Stordal, A. S., H. A. Karlsen, G. Nævdal, H. J. Skaug, and B. Vallès, 2011: Bridging the ensemble kalman filter and particle filters: the adaptive gaussian mixture filter. *Computational Geosciences*, **15**, 293–305.

Accepted for publication in *Monthly Weather Review.* DOI 10.1175/MWR-D-22-0108.1.

Thadewald, T., and H. Büning, 2007: Jarque–bera test and its competitors for testing normality – a power comparison. *Journal of Applied Statistics*, **34 (1)**, 87–105.

Villaseñor, J., and E. González-Estrada, 2009: A generalization of shapiro–wilk's test for multivariate normality. *Communications in Statistics—Theory and Methods*, **38**, 1870–1883, doi: 10.1080/03610920802474465.

Vukicevic, T., T. Greenwald, M. Zupanski, D. Zupanski, T. V. Haar, and A. S. Jones, 2004: Mesoscale cloud state estimation from visible and infrared satellite radiances. *Mon. Wea. Rev.*, **132**, 3066–3077.

Whitaker, J. S., and T. M. Hamill, 2002: Ensemble data assimilation without perturbed observations. *Mon. Wea. Rev.*, **130**, 1913–1924.

Zhang, F., C. Snyder, and J. Sun, 2004: Impacts of initial estimate and observation availability on convective-scale data assimilation with an ensemble Kalman filter. *Mon. Wea. Rev.*, **132**, 1238–1253.

Accepted for publication in *Monthly Weather Review.* DOI 10.1175/MWR-D-22-0108.1.