Volume 230, Nos. 1–2, June 2007 ISSN 0167-2789

PHYSICA D
NONLINEAR PHENOMENA

Data Assimilation

Guest Editors:
Kayo Ide
Christopher K.R.T. Jones

complete volume

Available online at
ScienceDirect
www.sciencedirect.com

http://www.elsevier.com/locate/physd

Available online at www.sciencedirect.com
ScienceDirect
Physica D 230 (2007) 127–145
PHYSICA D
www.elsevier.com/locate/physd

Data assimilation by field alignment

Sai Ravela*, Kerry Emanuel, Dennis McLaughlin

54-1624, Massachusetts Institute of Technology, Cambridge, MA 02139, United States

Available online 28 November 2006

Abstract

Classical formulations of data assimilation, whether sequential, ensemble-based or variational, are amplitude adjustment methods. Such approaches can perform poorly when forecast locations of weather systems are displaced from their observations. Compensating position errors by adjusting amplitudes can produce unacceptably "distorted" states, adversely affecting analysis, verification and subsequent forecasts.

There are many sources of position error. It is non-trivial to decompose position error into constituent sources and yet correcting position errors during assimilation can be essential for operationally predicting strong, localized weather events such as tropical cyclones.

In this paper, we propose a method that accounts for both position and amplitude errors. The proposed method assimilates observations in two steps. The first step is *field alignment*, where the current model state is aligned with observations by adjusting a continuous field of local displacements, subject to certain constraints. The second step is amplitude adjustment, where contemporary assimilation approaches are used. We demonstrate with 1D and 2D examples how applying field alignment produces better analyses with sparse and uncertain observations.

Keywords: Data assimilation; Variational methods; Ensemble filters; Alignment models; Position errors; Nonlinear geosciences; Computational science

1. Introduction

It is important to accurately forecast the positions of localized weather phenomena such as thunderstorms, squall lines, hurricanes, precipitation, and fronts. Failure to do so can result in a significant cost to commerce and life. Position errors occur frequently in forecasts and, as noted by several researchers, are a concern. For example, Alexander et al. [1] have shown that mesoscale simulations of marine cyclones using the MM5v1 [2] model initialized with NCEP operational analyses consistently have position errors in real forecast experiments. Thiebaux et al. [3] have shown that position errors of major forecast features were negatively affecting quality control decisions in NCEP models, and the latter were improved when position errors were removed. Sensitivity studies of precipitation data conducted by Jones and MacPherson [4] show that position errors lead to significant degradation of forecasts.

Addressing the causes of position errors directly is difficult. Hoffman et al. [5,6] argue that timing errors frequently generate position errors. Mariano [7] argues that inadequate data and model resolution, incorrect model physics, or simplifying dynamical/statistical assumptions lead to position errors. To be sure, position errors can arise from errors in background flow, model parameters, inadequate model resolution, existence of multi-scale interactions, and approximations of governing equations and parameterizations, among others. If there were a few parameters that could be estimated to resolve position errors, then a parameter estimation method could be used. The problem is that decomposing position errors into a few constituent sources is non-trivial and yet position errors cannot be ignored as operational examples suggest.

Indeed, this problem is significant enough that operational schemes have been developed. For example, tropical cyclones are often forecast to be in the wrong position, with ill-defined centers. In "bogussing", a technique employed operationally to overcome this difficulty, a bogus cyclone with an assumed structure and location/intensity is artificially introduced into the analysis cycle either in the first guess or as synthetic observations. The addition of the bogus vortex must be complemented with the removal of the vortex in the forecast.

* Corresponding address: Massachusetts Institute of Technology, Earth, Atmospheric and Planetary Science, 48-208, 15 Vassar Street, Cambridge, MA 02139, United States. Tel.: +1 617 2531969.
E-mail address: ravela@mit.edu (S. Ravela).

0167-2789/$ - see front matter
doi:10.1016/j.physd.2006.09.035

"Addition" and "deletion" must be conducted with care. If the addition is not "smooth", significant shocks can be introduced to the system and produce poor forecasts. If the incorrect vortex is not removed properly, the resulting "ghost" vortex can severely degrade the track forecast [8].

Bogussing attempts to correct a problem that contemporary data assimilation is not designed to handle. Current data assimilation approaches adjust amplitudes. In well defined storms, even small position errors will cause the first guess used in data assimilation to be spectacularly in error (in strength or amplitude) at grid points near both the real and first-guess storm positions. Thus, huge changes in first-guess amplitude may be necessary to produce an acceptable analysis. Brewster [9] argues that infrequent spatio-temporal features, incomplete data at small scales and a lack of background error estimates makes this problem hard to solve. Constructing background statistics that can optimally support large amplitude adjustments is problematic because it is difficult to construct amplitude statistics that model position statistics well. As we shall see, background error statistics that would be perfect without the existence of position errors will become inappropriate because position errors introduce bias (forecast mean and truth are displaced) and/or inflate the background error (co)variance. The effect of forecast bias has been examined before [10] and the effect of poor background error covariances on analyses are also known. As a result, a classical assimilation method (e.g. three-dimensional variational assimilation, or 3DVAR) will not always remove the feature from the incorrect location and rebuild it at the proper place suggested by the observations, but rather end up distorting the first guess.

Distorted analyses are a problem in their own right. An analysis may be declared to be optimal under some objective and yet bear little fidelity to what we expect to see. Distorted analyses are also a problem for generating subsequent forecasts. Thus, evidence from operational practice suggests that position errors must be corrected to produce good forecasts. Addressing modeling issues that lead to position errors is complicated and correcting position errors with contemporary assimilation methods can lead to distortions. In practice, even ad-hoc procedures seem to improve matters.

The starting point of this paper is the search for an objective method that can compensate for position errors in the assimilation cycle. We develop a formulation of the assimilation problem that adjusts both positions and amplitudes. Our formulation arises from a Bayesian view that conditions inference of displacements and amplitudes on certain priors. These priors are constructed from forecasts to capture amplitude statistics and *regularization* to specify displacement constraints. Together, the data likelihood and priors produce an optimization objective whose solution minimizes position and amplitude errors.

We show that, for large-scale problems, this objective can be solved in two steps. In the first step, called *field alignment*, model fields are aligned with observations. The alignment is carried out as a smooth (diffeomorphic) automatic remapping of the coordinate system underlying the state. The field alignment step essentially solves an auxiliary variational procedure that minimizes the misfit between forecast states and observations in the displacement space. Once fields are aligned, the second step, amplitude adjustment, can be carried out using contemporary data assimilation schemes. Thus, we synthesize a scheme where the newly proposed alignment step can be implemented as a *preprocessor* for both 3DVAR and ensemble approaches. We demonstrate, using 2D examples that the proposed methodology can significantly reduce the detrimental effects of position errors.

The remainder of this paper is organized as follows. In Section 2 related work is surveyed. In Section 3 the impact of position errors on data assimilation is examined. Then, in Section 4 we formulate the data assimilation problem in a manner that accounts for position errors. We then provide an efficient sequential solution in Section 5 and demonstrate its use on two-dimensional examples in Section 6. We conclude with a discussion of the relative merits of the proposed approach in Section 7.

2. Related work

Several solutions to the position error problem have been formulated. One solution is to *detect* features and formulate a position control problem using detected features [11]. In order to do this automatically, however, features need to be defined, detected, and correspondences between model-state and observations need to be established, automatically. Years of computer vision research has shown that these problems are extremely hard to solve. Features are not always readily detected either because they are not well-defined or because the data density is sparse, and the correspondence problem is NP-complete and thus lacks a deterministic polynomial time solution. Further, the alignment of features alone does not provide a framework for aligning the rest of the state unless additional constraints are imposed on the deformation.

Alexander et al. [1] propose a technique to improve the forecasts of feature locations associated with marine cyclones by using microwave integrated water-vapor imagery (IWV) and image warping of forecast mesoscale fields. The warping technique is based on *manually* selecting corresponding features, solving for a third-degree polynomial transform relating corresponding point coordinates and warping the entire field using this transform. Although such warping techniques have been used in the digital image processing community, neither the physical basis that motivates their choice, nor how the technique can be extended for automatically generating an objective analysis is clear.

Mariano [7] developed a technique for "melding" fields by detecting contours and determining the correspondence automatically. The estimated field is a weighted combination of contours, which does not distort the analysis. Such an approach is difficult to use when fields with differing contrasts are compared, or when observations are sparse.

In contrast to work by Mariano [7] and Alexander et al. [1], the method proposed here employs no feature detection or correspondence. The amplitude error between the fields is used directly to synthesize the deformation (or warping).

In a recent paper, Brewster [9] develops an alignment algorithm using a displacement-based cost function and a penalty term based on the inverse of a second order autoregressive term. This method divides the domain into overlapping control volumes, and the displacement is computed in each volume. The size of the control volume is variable and a multi-resolution formulation is proposed. This technique is demonstrated on thunderstorm simulations. Although our technique is also based on an Eulerian formulation, it is strictly different. There is no use of an autoregressive constraint and our result generalizes to diffeomorphic position errors.

In two papers, Hoffman et al. [5,6] develop a variational technique for producing analyses having displacement and amplitude errors, called distortion errors. Their formulation uses an objective with three terms in a spectral representation: a distortion cost function, a smoothness term for the distortion, and a barrier penalty term that bounds the distortion. The smoothness term is *global* and determined by a distortion (co)variance term. They demonstrate that using their method significantly reduces errors between ECMWF analyses and microwave IWV observations. In contrast, we use a *local* constraint for relating displacements. In their technique, the covariance term is designed to capture the error correlations between positions and amplitudes. We think that the correlations between amplitude errors and displacement errors are non-trivial and propose a solution that does not require it. Further, we do not use a barrier penalty term and our implementation is spatial.

3. Data assimilation with position errors

To compensate for nonlinearity arising from position errors, we represent the misfits between model and data phenomenologically, as errors in position and amplitude. This leads to a new formulation of the assimilation objective. Before we describe the details of this method, it is instructive to understand the complications arising from position errors in current assimilation methods.

Consider a Bayesian formulation of the data assimilation problem [12]. The state vector X_n at a discrete time t_n can be estimated using all measurements $Y_{0:n}$ from an associated conditional probability density $P(X_n|Y_{0:n})$. If we suppose that the dynamics is Markov in time, that is $P(X_n|X_{0:n-1}) = P(X_n|X_{n-1})$ and the observation errors are uncorrelated in time, and therefore $P(Y_{0:n}) = \prod_{i=0}^{n} P(Y_i)$, then $P(X_n|Y_{0:n})$ can be evaluated via Bayes' rule as:

$$P(X_n|Y_{0:n}) \propto P(Y_n|X_n)P(X_n|Y_{0:n-1}) \tag{1}$$

$$= P(Y_n|X_n)\int P(X_n|X_{n-1}) \times P(X_{n-1}|Y_{0:n-1})\mathrm{d}X_{n-1} \tag{2}$$

$$P(X_n|Y_{0:n}) \propto P(Y_n|X_n)P(X_n^f) \tag{3}$$

Eq. (1) presents the Bayesian form, ignoring the normalizing constant $P(Y_n)$. Eq. (2) depicts the recursive formulation for filtering. The distribution at the previous time $P(X_{n-1}|Y_{0:n-1})$ can be used to construct a conditional prior denoted $P(X_n^f)$ in Eq. (3) and is called the forecast distribution at time t_n. Constructing this distribution is non-trivial because only in very simple cases is it possible to integrate Eq. (2) analytically. Modeling the forecast distribution remains one of the outstanding challenges of contemporary data assimilation, with approaches ranging from carefully crafted error uncertainties to Monte Carlo methods.

To simplify the ensuing discussion, we adopt a linear observation model, $Y_n = HX_n + \eta$, where η is additive measurement noise and uncorrelated in space and time, and H is a linear observation operator. Further, we drop the explicit dependence on time by writing $X = X_n$ and observation $Y = Y_n$. Then, if the distributions in question are assumed Gaussian, with $P(X^f) \sim N(X, B)$ and $\eta \sim N(0, R)$, the mean of $P(X_n|Y_{0:n})$ is equal to its mode, which is the value of X that minimizes the following quadratic objective, written with its solution:

$$J(X) = (X - X^f)^{\mathrm{T}}B^{-1}(X - X^f) + (Y - HX)^{\mathrm{T}}R^{-1}(Y - HX) \tag{4}$$

$$X = X^f + BH^{\mathrm{T}}(HBH^{\mathrm{T}} + R)^{-1}(Y - HX^f) \tag{5}$$

Eqs. (4) and (5) can be interpreted deterministically, where Y and X^f are fixed vectors called the observation and first guess, and X is the estimated state. The matrices B and R are the respective uncertainties in state and observations. As shown, these equations are a simplified version (due to the linear observation operator) of what is commonly known as "3DVAR" in the meteorological community [13–16].

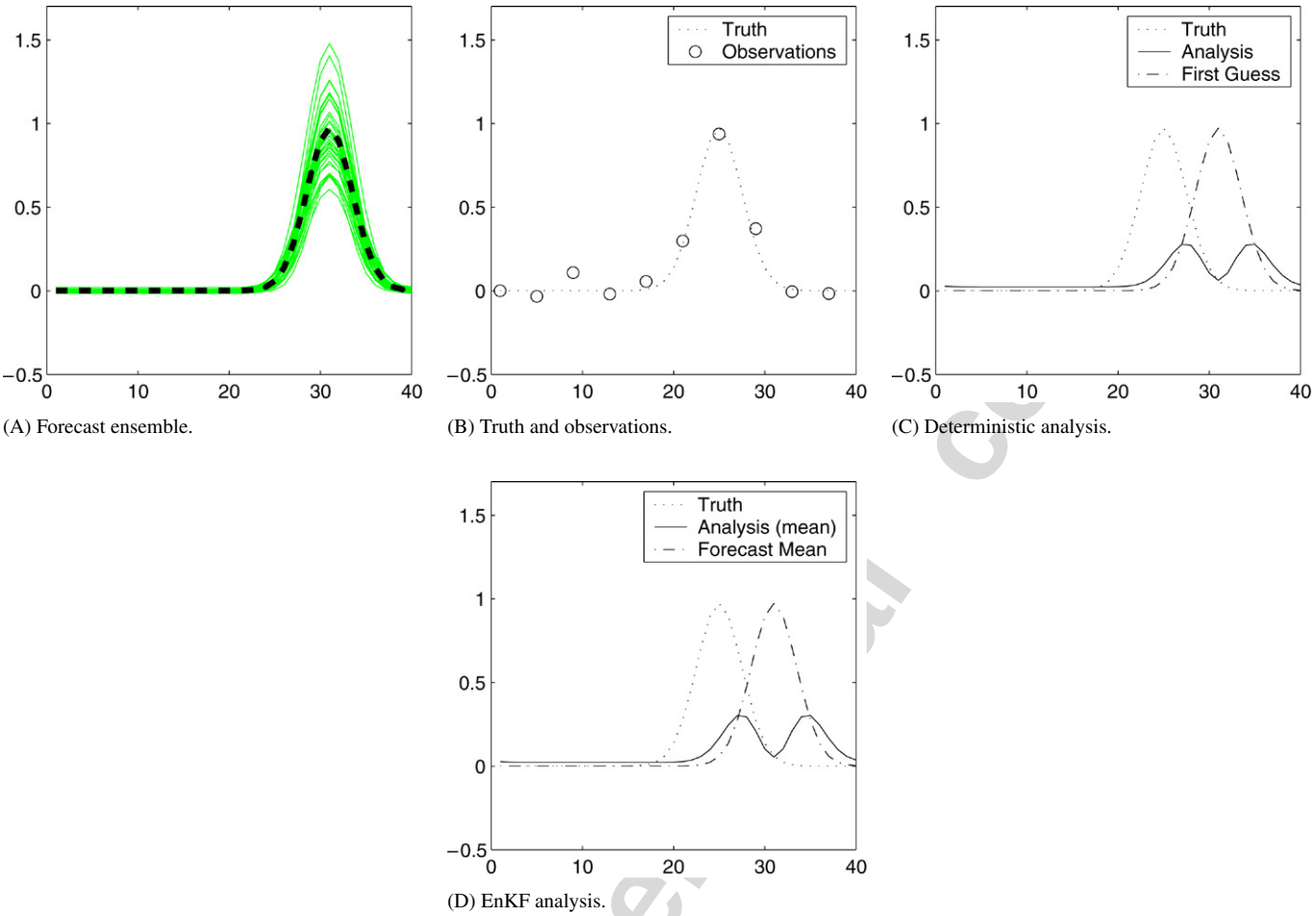
Eqs. (4) and (5) can also be interpreted probabilistically, which forms the basis for the Ensemble Kalman Filter [17]. In this case, an ensemble of estimates at time t_{n-1} are forecast to time t_n using the model. Let us call the forecast ensemble $A^f = [X_1^f \ldots X_S^f]$, where the columns of A^f are the S replicates of the ensemble. To implement the Ensemble Kalman Filter, we will assume that the observation equation is linear (as before) with spatially and temporally uncorrelated additive noise. We let Z represent a matrix of perturbed observations, $\tilde{A}^f$ be the deviation from mean $\bar{A}^f$ of A^f, the innovation covariance be $C = (H\tilde{A}^f)(H\tilde{A}^f)^{\mathrm{T}} + R$, and write:

$$A = A^f + \tilde{A}^f(H\tilde{A}^f)^{\mathrm{T}} \times [(H\tilde{A}^f)(H\tilde{A}^f)^{\mathrm{T}} + R]^{-1}(Z - HA^f) \tag{6}$$

$$= A^f + \tilde{A}^f(H\tilde{A}^f)^{\mathrm{T}}C^{-1}(Z - HA^f) \tag{7}$$

Both these methods perform equally poorly in the presence of position errors as the following examples will demonstrate.

Example 1. In Fig. 1, we show a forecast ensemble of one-dimensional fronts on a circular domain containing 40 nodes. The ensemble size is also 40. All the ensemble members have varying amplitudes, but exactly the same front positions. In this example, a front is defined as $\bar{X}^f(p) = \mathrm{e}^{-\frac{(p-p_0)^{\mathrm{T}}(p-p_0)}{2\sigma_0^2}}$ with $p_0 = 31$ and $\sigma_0 = 2$. An ensemble member X_i^f is generated by perturbing amplitude $X_i^f(p) = a_i * \bar{X}^f(p) + b_i$, where the scalars are random variables $a_i \sim N(1, 0.2^2)$ and



(A) Forecast ensemble. (B) Truth and observations. (C) Deterministic analysis.

(D) EnKF analysis.

Fig. 1. Panel (A) depicts a forecast ensemble with only amplitude perturbations. The ensemble mean is depicted as the dashed line. Panel (B) depicts the truth (dotted line) and observations. Panel (C) depicts a deterministic analysis (solid line) and Panel (D) depicts the analysis (mean) obtained from an ensemble Kalman filter (solid line).

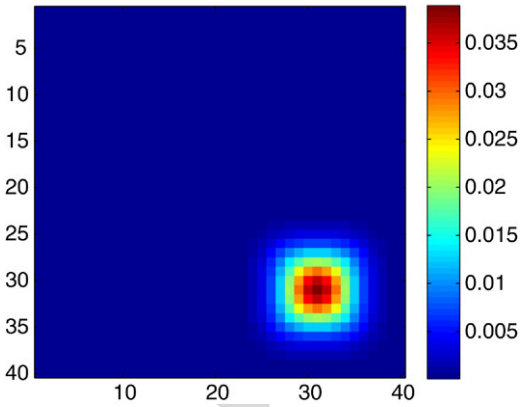


Fig. 2. The covariance computed from the forecast ensemble with only amplitude perturbations shown in Fig. 1(A).

$b_i \sim N(0, 0.01^2)$. The mean of the ensemble can be verified to be $\bar{X}^f(p)$. Fig. 1(A) shows the replicates and their mean (dashed line). Fig. 2 shows the covariance computed of the forecast ensemble.

Fig. 1(B) shows the "true" front as a dotted line. The truth X^{T} has exactly the same amplitude as the forecast ensemble mean $\bar{X}^f(p)$, except that it is displaced. The truth has a peak amplitude of 1 non-dimensional unit and smallest amplitude of zero.

Observations are generated by sampling the truth fairly densely in space (10 observations at every fourth node on a 40-dimensional state vector). Thus the observation operator H is a binary incidence matrix of size 10×40. The observational uncertainty is Gaussian, identical and independently distributed (iid) with a standard deviation of 0.05. This is substantially less than the background error variance at the front (see Fig. 2). Thus, the ith observation vector sample Y_i is synthesized from truth according to the linear observation equation $Y_i = HX^{\mathrm{T}} + \eta$, where $\eta \sim N(0, 0.05^2\mathbf{I})$. Note that, in this example, the forecast ensemble has Gaussian statistics, but there is a bias between truth and the forecast ensemble mean.

We implement a deterministic scheme, using as first guess the forecast mean $\bar{X}^f$. We use the forecast replicates to compute B statistically (see Fig. 2) and specify R and H from the above discussed construction. We then synthesize a single observation vector Y using the observation equation and then compute the innovation $\delta = Y - H\bar{X}^f$, and solve the linear system $\delta = (HBH^{\mathrm{T}} + R)\mu$ for the unknown vector μ using a conjugate gradient method (readily available in MATLAB as the *pcg* command). This is similar to the PSAS scheme [13],

with the key difference being the statistical computation of the background error covariance. Finally, we will obtain the solution as $\bar{X}^f + BH^{\mathrm{T}}\mu$.

In Fig. 1(C), the analysis from the procedure just discussed is shown. It is clear that the analysis (solid line) looks like neither the forecast ensemble nor the truth. It's somewhere in between, being pulled by the observations in some places and the background at others. It has replaced a single front with a bimodal front of far weaker strength.

We then implement an EnKF scheme. To do so, we compute the innovation covariance $C = HBH^{\mathrm{T}} + R$ by computing B from the ensemble and specifying R and H as in the deterministic case. We then compute the pseudo-inverse of C using singular value decomposition. Then, each forecast ensemble member $A^f(:, i) = X_i^f$ is paired with a perturbed observation $Z(:, i) = Y_i$ and applied to Eq. (6).[1] Fig. 1(D) shows the result of the ensemble Kalman filter, where the analysis is the ensemble mean. It is also distorted, although in a slightly different way.

Example 2. In Fig. 3, a forecast ensemble containing both amplitude and position error is generated. In this example, an ensemble member X_i^f is defined as $X_i^f(p) = a_i * \mathrm{e}^{-\frac{(p-p_i)^{\mathrm{T}}(p-p_i)}{2\sigma_0^2}} + b_i$, where σ_0, a_i and b_i and p_0 are defined in the same way as Example 1, and $p_i \sim N(p_0, \sigma_0)$. We then compute B from 40 ensemble members. The replicates and their mean are shown in Fig. 3(A). Fig. 3(B) depicts truth and observations. The matrices R and H defining the observation equation are exactly the same as in Example 1.

Fig. 3(C) depicts the result of the deterministic scheme (as discussed in Example 1), when the forecast mean is used as the first guess. It may be argued, in this case, that the ensemble mean is really not a good depiction of the front due to the large position errors, and hence a poor choice as a first guess. Therefore, we select as first guess a state X_i^f with $a_i = 1$, $b_i = 0$ and $p_i = p_0$. The result of doing so is shown in Fig. 3(D), with the analysis again being the solid line. In fact, any replicate in the forecast ensemble produces similarly distorted analysis. This can be seen from the Ensemble Kalman Filter (EnKF) solution. Fig. 3(E) depicts the performance of EnKF with the analysis-ensemble mean as the issued analysis. In each of these figures the analysis is also distorted.

Example 3. In Figs. 4 and 5, we construct a third example, with variability only in position. Fig. 4(A) depicts a forecast front. This front is an (unnormalized) Gaussian, of standard deviation 2.5 position units and is centered at a position of 25. That is, $X^f(p) = \mathrm{e}^{-\frac{(p-p_0)^{\mathrm{T}}(p-p_0)}{2\sigma_0^2}}$, where $p_0 = 25$ and $\sigma_0 = 2.5$. This forecast is used to produce a background covariance in the following way. We suppose that the front's position is uncertain and generate a large number (3000) samples by perturbing position. That is, we generate samples $X_i^f(p) = \mathrm{e}^{-\frac{(p-p_i)^{\mathrm{T}}(p-p_i)}{2\sigma_0^2}}$, where $p_i \sim N(p_0, \sigma_0)$. We then compute B from these samples.

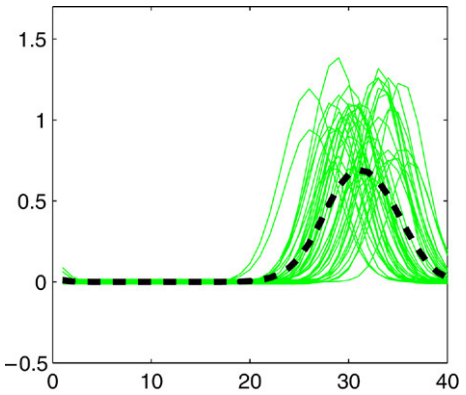
A series of "truths" are generated as $X_j^{\mathrm{T}}(p) = \mathrm{e}^{-\frac{(p-p_j)^{\mathrm{T}}(p-p_j)}{2\sigma_0^2}}$, where p_j is displaced farther away from the forecast position $p_0 = 25$, as shown in the left column of Fig. 5. In each instance, the truth is sampled every fourth location, for a total of 10 observations, to which Gaussian noise (iid) of standard deviation 0.05 is added. That is, $Y_j = X_j^{\mathrm{T}} + \eta$, where $\eta \sim N(0, 0.05^2\mathbf{I})$. By this construction, H and R are specified.

We then produce a deterministic analysis using X^f, B, H, R and Y_j as shown in the middle column of Fig. 5. These panels depict the truth (dotted line), the first guess (dash-dot line), which is the forecast, and an analysis (solid line). The analysis produced is thus 3DVAR (PSAS variation), using exactly the same scheme as discussed in Example 1 and also used for the deterministic analyses in Example 2. It can be seen that when the truth is close to the forecast, the analysis is quite good. In fact, it is quite good so long as the position error is smaller than σ_0. As it gets farther away, the background error covariance is not a good representation of the uncertainty and the analysis becomes distorted. Although this example suggests that introducing correlations across space (such as by perturbing position) can account for position errors, we will shortly see that this is not always a good choice.

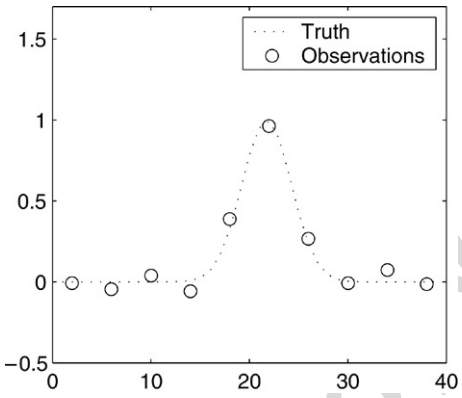
It is tempting to overcome this distortion with an isotropic background error covariance. The rationale is that an isotropic background error covariance could reflect the loss of flow dependence due to position uncertainty better than position perturbations. So we produce a deterministic analysis using an isotropic covariance B_{iso}. We compute the isotropic covariance as a circulant matrix constructed from a one-dimensional unnormalized Gaussian kernel. The peak amplitude of this Gaussian kernel is scaled to the maximum variance of the flow-dependent error covariance shown in Fig. 4(B), and is about 0.1. The standard deviation of this Gaussian is determined from the forecast shown in Fig. 4(A), by assuming a correlation length equal to the distance where the power of the forecast amplitude becomes half of its peak power. After some simple algebraic manipulation, we can determine the scale of the Gaussian kernel, σ_{iso}, used to produce the isotropic covariance as $\sigma_{\mathrm{iso}} = \frac{\sqrt{2\ln(2)}}{2}\sigma_0$. Since the forecast was generated with a Gaussian of standard deviation $\sigma_0 = 2.5$, therefore $\sigma_{\mathrm{iso}} = 1.47$. The resulting covariance B_{iso} is shown in Fig. 4(C).

We then produce a deterministic analysis using X^f, B_{iso}, H, R and Y_j as shown in the right column of Fig. 5. These panels depict the truth (dotted line), the first guess (dash-dot line), which is the forecast, and an analysis (solid line). The analysis produced is thus 3DVAR (PSAS variation), using exactly the same scheme as discussed in Example 1, the deterministic analyses in Example 2 and the flow-dependent case just discussed. Except for the forecast error covariance all other variables are exactly the same as in the flow-dependent case shown in the middle column of Fig. 5. It can be seen from

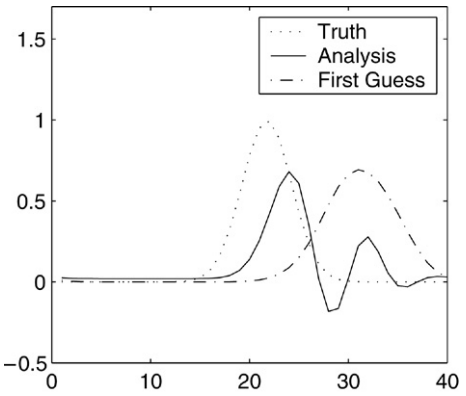
[1] In this case, the observations shown in Fig. 1(B) represents one perturbation.



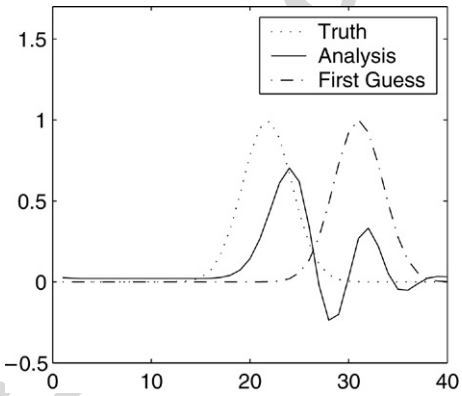
(A) Forecast ensemble.



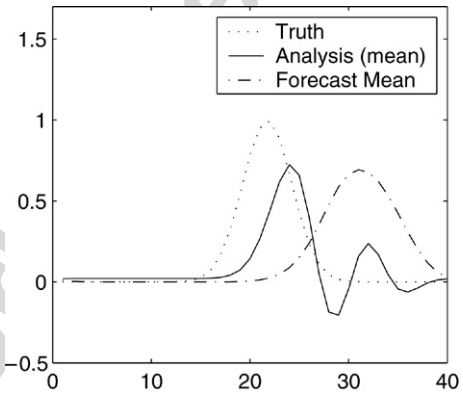
(B) Truth and observations.



(C) Deterministic analysis – 1.



(D) Deterministic analysis – 2.



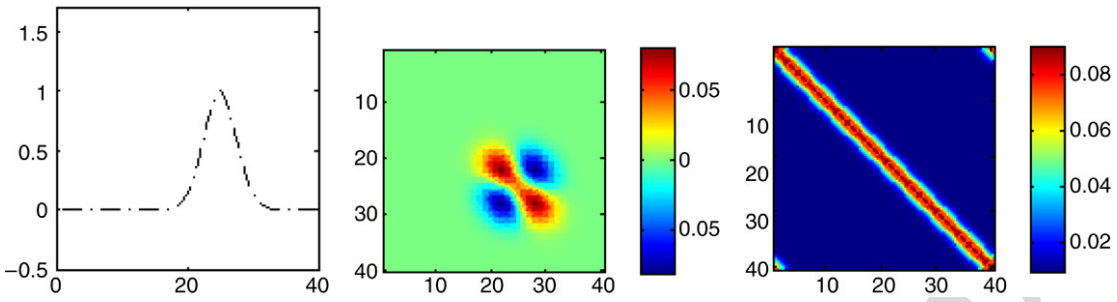
(E) EnKF analysis.

Fig. 3. Panel (A) shows the forecast ensemble containing amplitude and position perturbations. The forecast-ensemble mean is the dashed line. Panel (B) depicts the truth (dotted line) and observations. Panel (C) shows a deterministic analysis (solid line) with the forecast ensemble mean as the first guess. Panel (D) depicts a deterministic analysis using an ensemble member as the first guess. Panel (E) depicts the analysis ensemble mean (solid line) using an ensemble Kalman filter.

the right column of Fig. 5 that when the truth is close to the forecast, the analysis is also quite good. As it gets farther away, the isotropic background error covariance need not be a good representation of the uncertainty and the analysis also becomes distorted.

The effect of position errors is not surprising if one examines how position errors can violate the assumptions driving the estimation. The analysis can get distorted when the forecast distribution has a bias, as Example 1 shows. This is because the two sources of information, the observations and forecasts, are supposed to be unbiased, that is their expectations are supposed to be identical. Under a position error this assumption can be violated and thus produce an imperfect estimate even if the forecast ensemble is itself perfectly Gaussian.

In Example 2, the situation also includes one of poor (co)variance. Here, individual forecasts of the ensemble are seen as reasonably well-defined fronts. However, the forecast uncertainty contains spurious correlations in amplitudes due to the displacements between forecast replicates. The amplitude correlations become artificially broad and, therefore, forecast



(A) Forecast. (B) Background error covariance. (C) Isotropic covariance.

Fig. 4. Illustration of Example 3 (see text). Panel (A) shows the forecast. Panel (B) shows a background covariance (in color) constructed from the forecast. Panel (C) shows an isotropic background error covariance, which is also used in the experiment. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

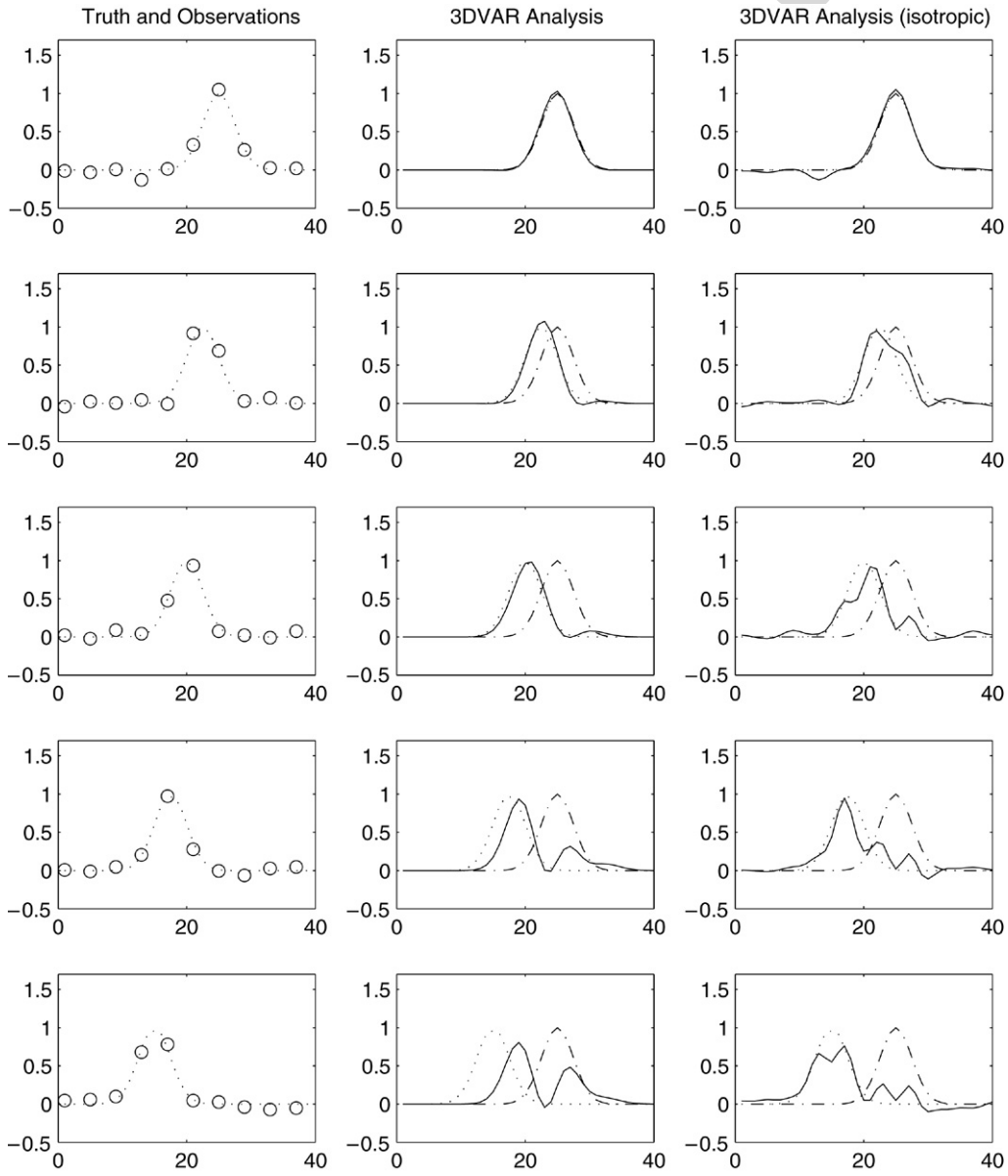


Fig. 5. Results for Example 3: The left column shows truth and observations. Panels down the middle column shows truth (dotted line), the first guess (dash-dot line) and the analysis (solid line) produced using 3DVAR with a flow-dependent covariance. The first guess is the same as the forecast in Fig. 4(A) and its uncertainty is depicted in Fig. 4(B). Panels down the right column shows truth (dotted line), the first guess (dash-dot line) and the analysis (solid line) produced using 3DVAR with an isotropic covariance. The first guess is the same as the forecast in Fig. 4(A) and its uncertainty is depicted in Fig. 4(C).

uncertainty becomes an inaccurate representation, owing to position error. As a result the analyses shown in Fig. 3(C–E), while closer to the truth in some areas, has also forced large distortions over a broader areas where observations are sparse or more uncertain than the background.

Problems can arise even in a deterministic setting such as Example 3. We designed a background error covariance to account for potential position errors by allowing position perturbations in the forecast. But doing so can only be effective to the degree that it does not inflate amplitude (co)variance artificially because this becomes much like the situation in Example 2 just discussed. Selecting an isotropic background error covariance also led to distortions. In the presence of sparse uncertain observations, new amplitudes are born at the right locations in the state but without necessarily removing the old ones to satisfaction.

The trade-off between position errors and amplitude covariance can be understood more formally. Suppose that we have forecasts X_i, and $i = 1 \ldots S$, which we deem a perfect Gaussian ensemble. Without loss of generality, we will represent this ensemble by a random variable X. We think of X as a spatially one-dimensional vector quantity, also referenced by a position variable p as $X(p)$. Let us suppose that $X \sim N(\bar{X}, C_{XX})$. That is, the covariance $C_{XX} = E_X[\tilde{X}\tilde{X}^{\mathrm{T}}]$, where E_X is the expectation under the distribution of X and $\tilde{X} = X - \bar{X}$. Let us consider a scalar position perturbation λ to this perfect forecast distribution, that is $X(p + \lambda)$. We now expand this via Taylor series (assuming the expansion is possible) as $X(p + \lambda) = X(p) + \lambda \frac{\partial X}{\partial p}$. The amplitude mean under the position perturbation can be written as $\bar{X}_\lambda = E_X[X(p + \lambda)] = \bar{X} + \lambda E_X\left[\frac{\partial X}{\partial p}\right]$. Further, we define the deviation of the gradients as $\Delta = \frac{\partial X}{\partial p} - E\left[\frac{\partial X}{\partial p}\right]$. Then the covariance as a result of the position perturbation is $C = C_{XX} + \lambda E_X[\tilde{X}\Delta^{\mathrm{T}}] + \lambda E_X[\Delta\tilde{X}^{\mathrm{T}}] + \lambda^2 E_X[\Delta\Delta^{\mathrm{T}}]$. If λ is assumed to be a Gaussian random variable with mean 0 and standard deviation σ_λ, then we can say that the expected amplitude covariance under the position perturbation is $E_\lambda[C] = C_{XX} + \sigma_\lambda^2 E_X[\Delta\Delta^{\mathrm{T}}]$. We define $C_\lambda = E_\lambda[C]$ and define $C_{\Delta\Delta} = E_X[\Delta\Delta^{\mathrm{T}}]$ and therefore, we have

$$C_\lambda = C_{XX} + \sigma_\lambda^2 C_{\Delta\Delta} \tag{8}$$

Eq. (8) says that the expected amplitude covariance as a result of random position perturbations is less certain or "inflated" than without. The degree of inflation depends, to the first order, on the spatial gradients and the variance of the position perturbation.

This has a bearing to both deterministic and ensemble assimilation schemes because, it is immaterial whether the forecast ensemble contains a large position spread (Example 2) or whether we allow artificial spatial correlations (Example 3) to account for potential position errors. In each case, the covariance is worse than the "perfect" one. Furthermore, even if the only position error were a large bias (Example 1), we still have a distortion problem. This situation is exacerbated when observations are sparse and uncertain because the imperfect background covariance can incorrectly adjust amplitudes. In an ensemble setting, if we suppose that forecasts arise from models, forecasting the mesoscale feature in the wrong place in various ways, ensemble amplitude-assimilation methods can no longer be expected to work well. In the deterministic framework, designing an appropriate background under (non-systematic) position errors becomes even more challenging. What we can do, as this paper describes, is to pose an objective that minimizes position and amplitude errors by using the observations and available forecast(s). This, we claim, compensates for position errors, even as it assimilates amplitude data.

4. Data assimilation by field alignment

To address the position error problem, we reformulate the classical quadratic objective in a way that allows position adjustments in addition to amplitude adjustments. The key step in this new approach is to explicitly represent and minimize position errors. Therefore, we introduce auxiliary control variables (displacements) that are estimated along with amplitudes. The displacement variables are defined at each node of the grid representing the state and specify a deformation of the grid. By using this scheme we can control both amplitudes and positions.

To make this framework more explicit it is useful to introduce some notation. Let $X = X(\mathbf{r}) = \{X[\underline{r}_1^{\mathrm{T}}] \ldots X[\underline{r}_m^{\mathrm{T}}]\}$ be the model-state vector defined over a spatially discretized computational grid Ω, and $\mathbf{r}^{\mathrm{T}} = \{\underline{r}_i = (x_i, y_i)^{\mathrm{T}}, i \in \Omega\}$ be the position indices. Similarly, let $\mathbf{q}$ be a *vector* of displacements. That is, $\mathbf{q}^{\mathrm{T}} = \{\underline{q}_i = (\Delta x_i, \Delta y_i)^{\mathrm{T}}, i \in \Omega\}$. Then the notation $X(\mathbf{r} - \mathbf{q})$ represents *displacement* of X by $\mathbf{q}$ (see Fig. 6). The displacement field $\mathbf{q}$ is real-valued, so $X(\mathbf{r} - \mathbf{q})$ must be evaluated by interpolation if necessary.

In a probabilistic sense, we may suppose that finding $(X, \mathbf{q})$ that has the maximum a posteriori probability in the distribution $P(X, \mathbf{q}|Y)$ is appropriate. Using Bayes' rule we obtain

$$P(X, \mathbf{q}|Y) \propto P(Y|X, \mathbf{q}) P(X^f|\mathbf{q}) P(\mathbf{q}) \tag{9}$$

As before, we assume a linear observation model with uncorrelated noise in space and time, and Markov dynamics. If we make a Gaussian assumption of the component densities, we can write:

(1) Data likelihood:

$$P(Y|X, \mathbf{q}) = \frac{1}{(2\pi)^{\frac{n}{2}} |R|^{\frac{1}{2}}} \times \mathrm{e}^{-\frac{1}{2}(Y - H\,X(\mathbf{r}-\mathbf{q}))^{\mathrm{T}} R^{-1} (Y - H\,X(\mathbf{r}-\mathbf{q}))} \tag{10}$$

This equation is the data-likelihood term. It implies that the observations can be related using a Gaussian model to the displaced state $X(\mathbf{r} - \mathbf{q})$, where $X(\mathbf{r})$ is defined on the original grid, and $\mathbf{q}$ is a displacement field. We use the linear observation model here, and therefore $Y = HX(\mathbf{r} - \mathbf{q}) + \eta$, $\eta \sim N(0, R)$. We should emphasize here that the observation vector is fixed. Its elements are always defined from the original grid.

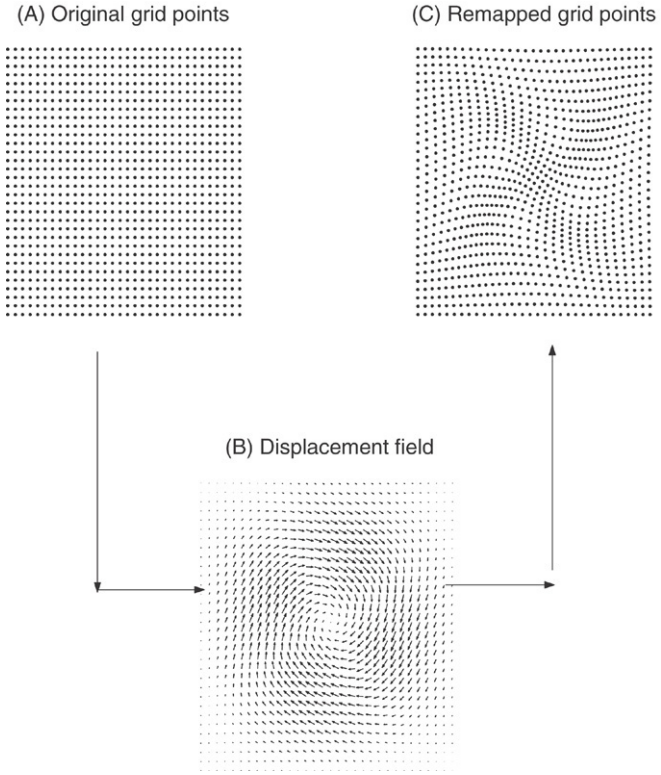


Fig. 6. A graphical illustration of field alignment. State vector on a discretized grid is moved by deforming its grid (**r**) by a displacement (**q**).

(2) Amplitude prior:

$$P(X^f|\mathbf{q}) = \frac{1}{(2\pi)^{\frac{n}{2}}|B(\mathbf{q})|^{\frac{1}{2}}} \times \mathrm{e}^{-\frac{1}{2}(X(\mathbf{r}-\mathbf{q})-X^f(\mathbf{r}-\mathbf{q}))^{\mathrm{T}}B(\mathbf{q})^{-1}(X(\mathbf{r}-\mathbf{q})-X^f(\mathbf{r}-\mathbf{q}))} \quad (11)$$

This equation defines the *amplitude prior*. Given a fixed displacement field **q** and a forecast $X^f(\mathbf{r})$ defined on the original grid, it states that the forecast distribution is assumed to be Gaussian in the position-corrected space, even if it isn't in the uncorrected space. Once we assume Gaussian statistics in a position-corrected space, it is immediately clear that the forecast statistics are conditioned on the displacement field. In particular, its second moment, the covariance B, is dependent on **q**. By simple analogy, if we associate an error covariance on the original forecast grid, this error covariance will have to be remapped when the forecast grid is deformed. Thus, we write the forecast covariance as $B(\mathbf{q})$.

(3) Displacement prior:

$$P(\mathbf{q}) = \frac{1}{\alpha}\mathrm{e}^{-L(\mathbf{q})} \quad (12)$$

This equation specifies a *displacement prior*. This prior is constructed from an energy function $L(\mathbf{q})$ which expresses constraints on the displacement field. The proposed method for constructing L is drawn from the nature of the expected displacement field. Displacements can be represented as smooth flow fields in many fluid flows and often arise from systematic and large-scale *background flow* errors, for example see [1]. Smoothness naturally leads to a Tikhonov type formulation [18] and, in particular, $L(\mathbf{q})$ is designed as a gradient and a divergence penalty term. These constraints, expressed in quadratic form, are:

$$L(\mathbf{q}) = \frac{w_1}{2}\sum_{j\in\Omega}\mathbf{tr}\{[\nabla \underline{q}_j][\nabla \underline{q}_j]^{\mathrm{T}}\} + \frac{w_2}{2}\sum_{j\in\Omega}[\nabla\cdot \underline{q}_j]^2 \quad (13)$$

In Eq. (13), $\mathbf{q}_j$ refers to the jth grid index and **tr** is the trace. Eq. (13) is a *weak constraint*, weighted by the corresponding weights w_1 and w_2. Note that the constant α can be defined to make Eq. (12) a proper probability density. In particular, define $Z(\mathbf{q}) = \mathrm{e}^{-L(\mathbf{q})}$ and define $\alpha = \int_{\mathbf{q}} Z(\mathbf{q})\mathrm{d}\mathbf{q}$. This integral exists and converges.

It is instructive to note that the constraints we impose are *local* and can be contrasted to other approaches that assume that the displacement field **q** has a particular distribution. The most tempting is to assume that **q** has a uniform distribution. Doing so, however, does not constrain the solution at all; the prior is uninformative and has little physical validity. Second, is to assume that **q** follows a Gaussian, or more strongly, it is jointly Gaussian with amplitude errors. This results in an approach similar to [5, 6], but it is unclear how to estimate the parameters of such a distribution [19]. It is precisely this lack of knowledge of the displacement prior that leads us to propose smoothness constraints which, as the preceding discussion shows, can be interpreted to have a Gaussian distribution.

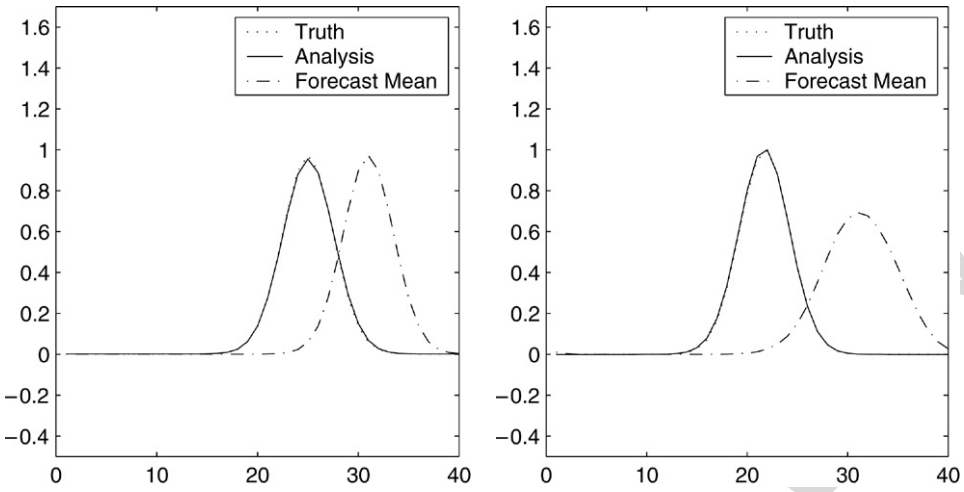
With these definitions of probabilities, we are in a position to construct an objective by evaluating the log probability. After defining $\mathbf{p} = \mathbf{r} - \mathbf{q}$, we can state an objective as:

$$J_2(X,\mathbf{q}) = \frac{1}{2}(X(\mathbf{p}) - X^f(\mathbf{p}))^{\mathrm{T}}B(\mathbf{q})^{-1}(X(\mathbf{p}) - X^f(\mathbf{p})) + \frac{1}{2}(Y - H\ X(\mathbf{p}))^{\mathrm{T}}R^{-1}(Y - H\ X(\mathbf{p})) + L(\mathbf{q}) - \frac{1}{2}\ln(|B(\mathbf{q})|) \quad (14)$$

Solving this objective is not easy. It isn't clear how to compute $B(\mathbf{q})$. Neither is it clear that the gradients can be computed easily. These difficulties can be overcome by making several approximations.

The first choice we make is to consider a statistical representation of uncertainty, using an *ensemble of forecast states*. In this case, computing B is straightforward. Let us suppose that S samples $\mathbf{X} = X_1 \ldots X_S$ are to be estimated along with associated displacements $Q = \mathbf{q_1}, \ldots, \mathbf{q_S}$, from S forecasts $X_s^f, s = 1 \ldots S$. Let $\mathbf{p_s} = \mathbf{r} - \mathbf{q_s}$ and $\bar{X}^f = \frac{1}{S}\sum_{s=1}^{S} X_s^f(\mathbf{p_s})$. The background error covariance is:

$$B_Q = B(\mathbf{X^f}; Q) = \frac{1}{S-1}\sum_{s=1}^{S} \times (X_s^f(\mathbf{p_s}) - \bar{X}^f)(X^f(\mathbf{p_s}) - \bar{X}^f)^{\mathrm{T}} \quad (15)$$



(A) One-step analysis (Example 1). (B) One-step analysis (Example 2).

Fig. 7. Results of the one-step method. Panel (A) depicts the analysis (solid line), truth (dotted line) and forecast ensemble mean (dash-dot line) obtained for Example 1 (Fig. 1). Panel (B) depicts the analysis, truth and forecast ensemble mean, for Example 2 (Fig. 3).

The ensemble framework leads to a modified objective, which we can write as:

$$J(\mathbf{X}, Q) = \frac{1}{S}\sum_{s=1}^{S} J_s(\mathbf{X}, Q) \tag{16}$$

where J_s is defined as, $for\ s = 1 \ldots S$

$$\begin{aligned} J_s(\mathbf{X}, Q) = {} & \frac{1}{2}(X_s(\mathbf{p_s}) - X_s^f(\mathbf{p_s}))^{\mathrm{T}} B_Q^{-1}(X_s(\mathbf{p_s}) - X_s^f(\mathbf{p_s})) \\ & + \frac{1}{2}(Y_s - H\ X_s(\mathbf{p_s}))^{\mathrm{T}} R^{-1}(Y_s - H\ X_s(\mathbf{p_s})) \\ & + L(\mathbf{q_s}) - \ln(|B_Q|) \end{aligned} \tag{17}$$

Eq. (17) can be used to align and estimate amplitudes for each ensemble member. Note that each ensemble member is paired with a perturbed observation Y_s, just as in the original EnKF framework.

The second choice we make is to construct an iterative solution of the inference problem. That is, we define

$$J(\mathbf{X}, Q|Q^-) = \frac{1}{S}\sum_{s=1}^{S} J_s(\mathbf{X}, Q|Q^-) \tag{18}$$

where $J_s(\mathbf{X}, Q|Q^-)$ is defined as:

$$\begin{aligned} J_s(\mathbf{X}, Q|Q^-) = {} & \frac{1}{2}(X_s(\mathbf{p_s}) - X_s^f(\mathbf{p_s}))^{\mathrm{T}} \\ & \times B_{Q^-}^{-1}(X_s(\mathbf{p_s}) - X_s^f(\mathbf{p_s})) \\ & + \frac{1}{2}(Y_s - H\ X_s(\mathbf{p_s}))^{\mathrm{T}} R^{-1}(Y_s - H\ X_s(\mathbf{p_s})) \\ & + L(\mathbf{q_s}) - \ln(|B_{Q^-}|) \end{aligned} \tag{19}$$

Here, B_{Q^-} is a notation that it is fixed during the evaluation of the objective. We therefore have an iterative basis for a solution that avoids computation of derivatives of higher-order terms. The gradients at an iteration can now be written for each ensemble member $s = 1 \ldots S$ as:

$$\begin{aligned} \frac{\partial J_s}{\partial \mathbf{q_s}} = {} & (\nabla X_s|_{\mathbf{p_s}} - \nabla X_s^f|_{\mathbf{p_s}})^{\mathrm{T}} B_{Q^-}^{-1}(X_s(\mathbf{p_s}) - X_s^f(\mathbf{p_s})) \\ & + \nabla X_s|_{\mathbf{p_s}} H^{\mathrm{T}} R^{-1}(H\ X_s(\mathbf{p_s}) - Y_s) + \frac{\partial L}{\partial \mathbf{q_s}} \end{aligned} \tag{20}$$

$$\begin{aligned} \frac{\partial J_s}{\partial X_s} = {} & B_{Q^-}^{-1}(X_s(\mathbf{p_s}) - X_s^f(\mathbf{p_s})) \\ & + H^{\mathrm{T}} R^{-1}(H\ X_s(\mathbf{p_s}) - Y_s) \end{aligned} \tag{21}$$

$$\frac{\partial L}{\partial \underline{q}_{s,i}} = w_1 \nabla^2 \underline{q}_{s,i} + w_2 \nabla(\nabla \cdot \underline{q}_{s,i}) \tag{22}$$

The term $\underline{q}_{s,i}$ is the displacement at a grid point i associated with an ensemble member s.

To minimize Eq. (18), the gradients are computed using Eqs. (20)–(22), and all of this is supplied to a standard optimization code. To be sure, the optimization proceeds with initialization $X_s = X_s^f$ and $Q = 0$. Every time the search algorithm requests an evaluation of the objective and its gradient, the aforementioned equations are used with a new set of perturbed observations. The search algorithm terminates when there is no improvement in objective or an iteration limit has been reached. We call this the *one-step* algorithm.

For the 1D example, Fig. 7(A), (B) demonstrate the estimate (solid line) produced using a quasi-Newton algorithm, with a BFGS[2] Hessian update scheme [20–23], available as part of the MATLAB implementation (medium-scale *fminunc* command). We have also tried a conjugate-gradients implementation, with no different results in these examples. In the cases shown in Fig. 7(A), (B), the analysis issued is the final ensemble mean.

There are several points worth mentioning about the one-step algorithm:

(1) Alignment is expressly Eulerian, individual features are neither identified nor required for alignment, although

[2] Broyden–Fletcher–Goldfarb–Shanno scheme.

featuredness or texture clearly influences the solution. The deformation is defined on the continuum and evolves over iterations.

(2) The constraint in L is modeled as such because we expect the fluid flow to be smooth. From a regularization point of view, there can be other choices [24] as well.

(3) The regularization constraint is a weak constraint. This implies that the constraints are not satisfied exactly. The weights determine how strongly the constraints influence the flow field.

The one-step algorithm essentially validates the idea of simultaneous alignment and amplitude assimilation, but there are several limitations.

First, this algorithm would scale poorly with state and ensemble size. At each iteration the background covariance, its inverse and determinant will have to be computed. This can become prohibitive, even if the approach shares in flavor the ensemble Kalman filter formulation for updating each ensemble member separately.

Second, for two reasons, this method cannot be used to solve the deterministic situation presented in Example 3. There is no easy way to reshape the background error covariance depicted in Fig. 4(B), as the preceding discussion has already highlighted. So the one-step scheme really cannot function in a purely deterministic mode. Even if we assume that this uncertainty can be statistically computed via an ensemble, in this particular example, the ensemble is purely position-perturbed and has no amplitude perturbations. If we suppose that each ensemble member is aligned perfectly, then the ensemble collapses, presenting a practical difficulty.

5. Sequential solution

We propose a variation of the solution that can function with deterministic or ensemble approaches alike, and can tackle the dimensionality problem better. Instead of solving the displacements and amplitudes "simultaneously", we make an approximation using the Euler–Lagrange equations. Following Eqs. (20) and (21), these can be written as:

$$\frac{\partial J_s}{\partial \mathbf{q_s}} = (\nabla X_s|_{\mathbf{p_s}} - \nabla X_s^f|_{\mathbf{p_s}})^{\mathrm{T}} B_{Q^-}^{-1}(X_s(\mathbf{p_s}) - X_s^f(\mathbf{p_s})) + \nabla X_s|_{\mathbf{p_s}} H^{\mathrm{T}} R^{-1}(H\ X_s(\mathbf{p_s}) - Y_s) + \frac{\partial L}{\partial \mathbf{q_s}} = 0 \quad (23)$$

$$\frac{\partial J_s}{\partial X_s} = B_{Q^-}^{-1}(X_s(\mathbf{p_s}) - X_s^f(\mathbf{p_s})) + H^{\mathrm{T}} R^{-1}(H\ X_s(\mathbf{p_s}) - Y_s) = 0 \quad (24)$$

It is clear that these equations are highly non-linear functions in X_s and $\mathbf{q_s}$. We solve them sequentially, in exactly two steps. In the first step we fix X_s to X_s^f in Eq. (23) and solve for $\hat{Q}$. We then use this solution to solve Eq. (24). The sequential approach is a *two-step* approximation. The first step is the displacement or alignment equation, written as:

$$\frac{\partial L}{\partial \mathbf{q_s}} + \nabla X_s^{fT}|_{\mathbf{p_s}} H^{\mathrm{T}} R^{-1}(H\ X_s^f(\mathbf{p_s}) - Y_s) = 0 \quad (25)$$

Using the regularization constraints the alignment equation at a node i now becomes:

$$w_1 \nabla^2 \underline{q}_{s,i} + w_2 \nabla(\nabla \cdot \underline{q}_{s,i}) + [\nabla X_s^{fT}|_{\mathbf{p_s}} H^{\mathrm{T}} R^{-1} \times (H[X_s^f(\mathbf{p_s})] - Y_s)]_i = 0 \quad (26)$$

This is the field alignment formulation where, phenomenologically, Eq. (26) introduces a forcing based on the residual between the model-field and observation-field modulated by the local gradient. The constraints on the displacement field allow the forcing to propagate to a consistent solution. Unfortunately, Eq. (26) is also non-linear, and is therefore solved iteratively, where it becomes a Poisson equation. During each iteration $\mathbf{q_s}$ is computed by holding the forcing term constant. The estimate of displacement at each iteration is then used to deform a copy of the original forecast model-field using bi-cubic interpolation. Together with a perturbed observation sample, the forcing for the next iteration is generated. The process is repeated until a small displacement residual is obtained, the misfit with observations does not improve, or an iteration limit is reached. Upon convergence, we have an aligned forecast ensemble $X_s^f(\hat{\mathbf{p}}_\mathbf{s})$, $s = 1 \ldots S$, from which $B_{\hat{Q}}$ is computed. With these quantities, the amplitude recovery is written as:

$$X_s(\hat{\mathbf{p}}_\mathbf{s}) = X_s^f(\hat{\mathbf{p}}_\mathbf{s}) + B_{\hat{Q}} H^{\mathrm{T}} \times (H B_{\hat{Q}} H^{\mathrm{T}} + R)^{-1}(Y_s - H\ X_s^f(\hat{\mathbf{p}}_\mathbf{s})) \quad (27)$$

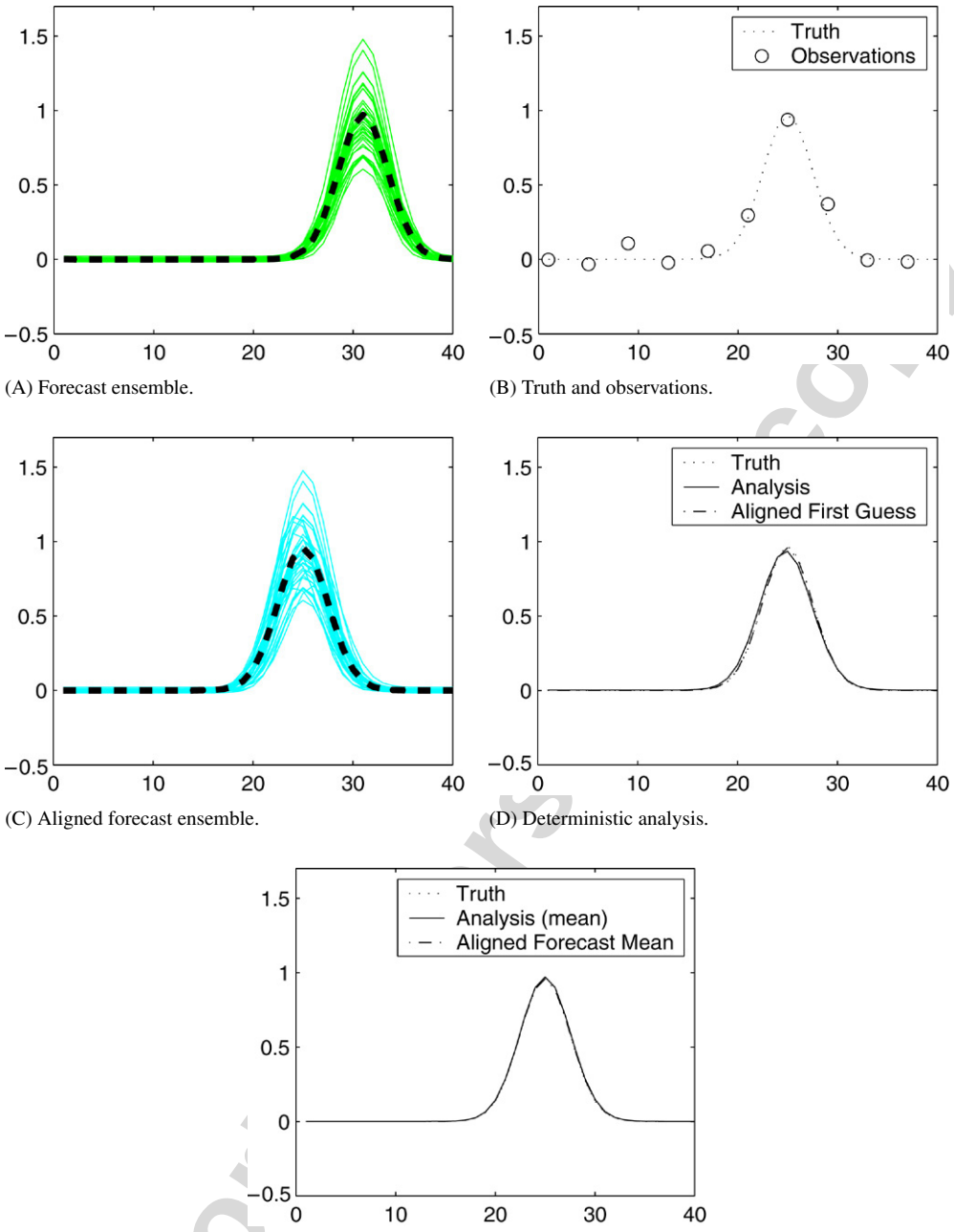
Eq. (27) can be implemented using several familiar schemes. Here we outline three:

Ensemble scheme: Eq. (27) is the basis for the ensemble Kalman filter, because $\hat{Q}$ is fixed. In the simplest interpretation, each position-corrected forecast replicate $X_s^f(\hat{\mathbf{p}}_\mathbf{s})$ can be paired with a perturbed observation Y_s, to produce a filtered estimate X_s. Thus, when the entire forecast ensemble is updated, a posterior distribution is obtained, from which the analysis can be issued (typically the mean). Of course, we must point out that alternative schemes such as the square-root formulations can be implemented because Eq. (27) expresses the standard quadratic update equation that forms the basis for many assimilation methods.

Deterministic ensemble scheme: In this mode, one can compute the background after aligning individual replicates. The aligned replicates can be used to compute a statistical background error covariance. Once this is done, deterministic assimilation proceeds from a single aligned first guess.

Purely deterministic scheme: The proposed two-step method can also be used when a forecast ensemble is not available. Since the alignment does not depend on the background error covariance, one can align the first-guess field and use the aligned first-guess to craft a state-dependent background error covariance [25]. In essence, the same procedure that would be used with the unaligned first-guess can be also used with the aligned first-guess.

These schemes are demonstrated on the one-dimensional examples used thus far. The matrices H and R are the same



(A) Forecast ensemble.
(B) Truth and observations.
(C) Aligned forecast ensemble.
(D) Deterministic analysis.
(E) EnKF analysis.

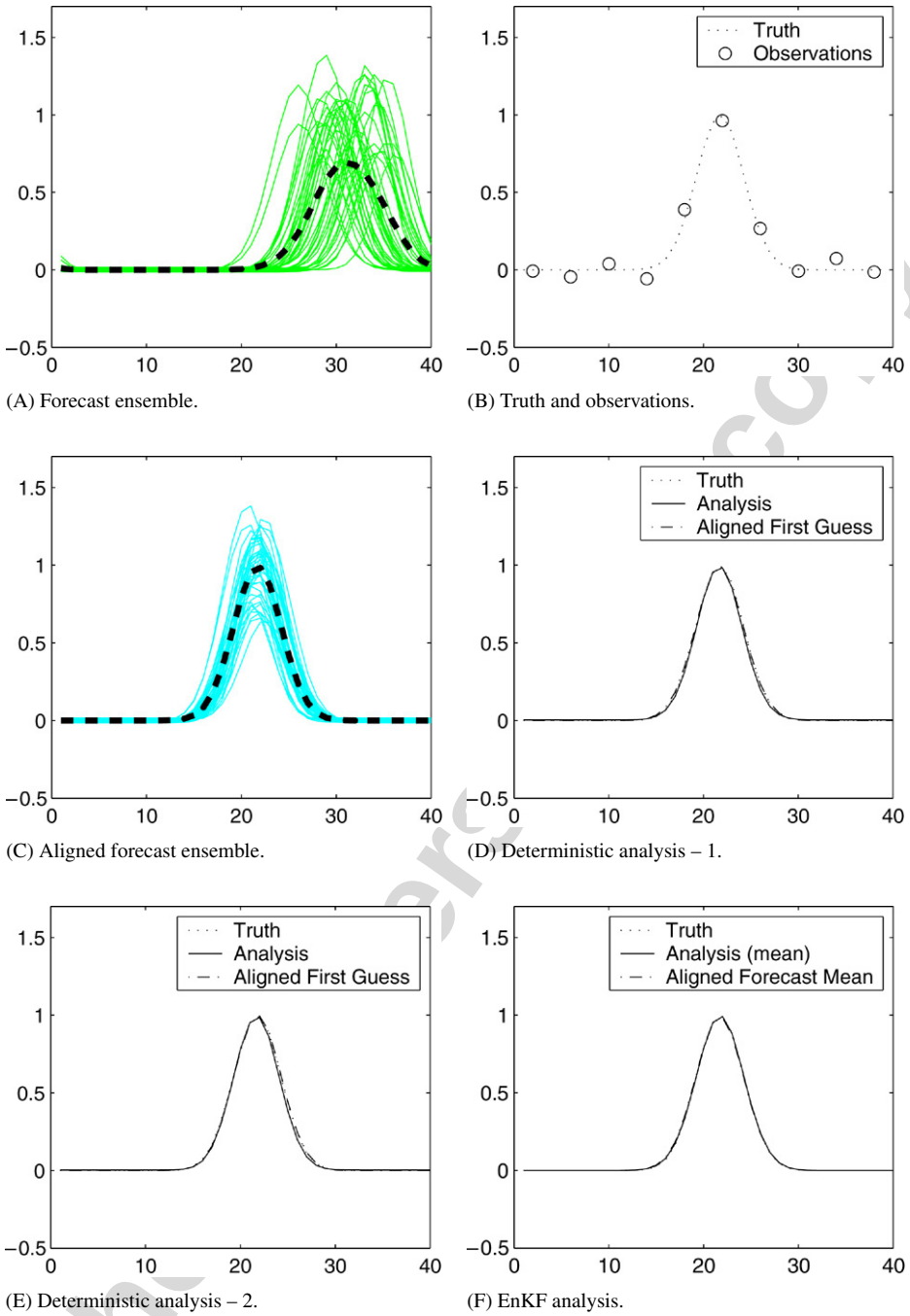
Fig. 8. Two-step solution for Example 1. Panel (A) depicts the original forecast ensemble and its mean. Panel (B) depicts the truth and observations, also the same as Fig. 1(B). Panel (C) depicts the aligned forecast ensemble and its mean (step 1). Panel (D) depicts the deterministic analysis after alignment (step 2). Panel (E) depicts the analysis using the Ensemble Kalman Filter (step 2) after alignment.

as used to compute the previous solutions. In deterministic schemes the observation vector will be the same, and in ensemble schemes the perturbed observation ensembles are also the same as the corresponding ones in the previous results. The difference is in the use of the aligned forecast (ensemble) to recompute the background and its covariance.

Fig. 8 shows the result of the two-step approach on Example 1. Panel (A) depicts the unaligned forecast ensemble and its mean, while panel (B) depicts truth and station observations. Panel (C) shows the result of alignment with the two-step approach, depicting the aligned ensemble and its mean. From this starting point both deterministic and ensemble assimilation schemes produce better analyses, as shown in panels (D) and (E).

Similarly, Fig. 9 demonstrates the improvements for the analogous situation depicted in Example 2 (Fig. 3). The deterministic analyses with the aligned forecast ensemble mean (panel D) or an ensemble member (panel E), or the EnKF solution (panel F), all demonstrate improved analyses. The two-step solution is also seen to be close to the one-step solution in these examples, though this need not be the case in general.

Finally, Fig. 10 depicts the result of applying the two-step solution to Example 3, specifically for the case shown in the bottom row of Fig. 5. We closely simulate a purely

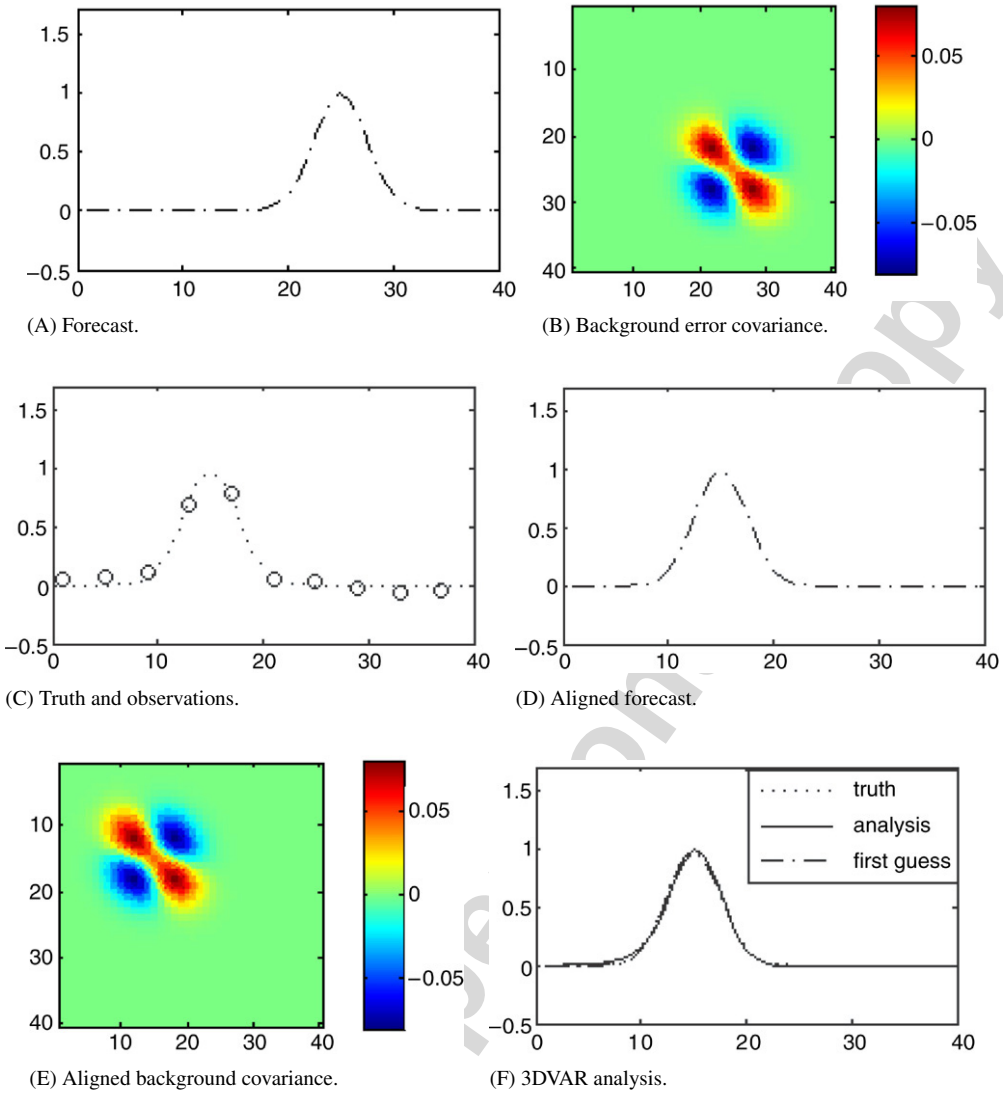


(A) Forecast ensemble.
(B) Truth and observations.
(C) Aligned forecast ensemble.
(D) Deterministic analysis – 1.
(E) Deterministic analysis – 2.
(F) EnKF analysis.

Fig. 9. Two-step solution for Example 2. Panel (A) depicts the original forecast ensemble and its mean. Panel (B) depicts the truth and observations, also the same as Fig. 3(B). Panel (C) shows the aligned forecast ensemble and its mean (step 1). Panel (D) depicts the deterministic analysis using the aligned ensemble mean as the first guess (step 2). Panel (E) depicts deterministic analysis using an ensemble member as the first guess (step 2), and panel (F) shows the analysis using the Ensemble Kalman Filter (step 2).

deterministic formulation, in the sense that only the forecast shown in Fig. 10(A) is aligned with the observations shown (along with the truth) in Fig. 10(C). This aligned forecast is shown in Fig. 10(D). The aligned forecast error covariance is generated using the same position perturbation scheme for the flow-dependent case as the unaligned version in Fig. 5. Thus, rather than use an ensemble scheme and align each ensemble member to generate the aligned forecast error covariance, we have aligned the forecast to produce an aligned first guess and then produce a new forecast error covariance. This is shown in Fig. 10(E) and can be contrasted with the unaligned error covariance shown in Fig. 10(B). As a result, the analysis using a 3DVAR scheme (same as in Examples 1–3), is much closer to the truth, shown as the solid line in Fig. 10(F).

Improvements in estimation can also be seen readily with an isotropic background error covariance. As shown in Fig. 11, the aligned forecast in panel A (same as Fig. 10(D)) is used with the isotropic background error covariance in panel B (same as



(A) Forecast. (B) Background error covariance.

(C) Truth and observations. (D) Aligned forecast.

(E) Aligned background covariance. (F) 3DVAR analysis.

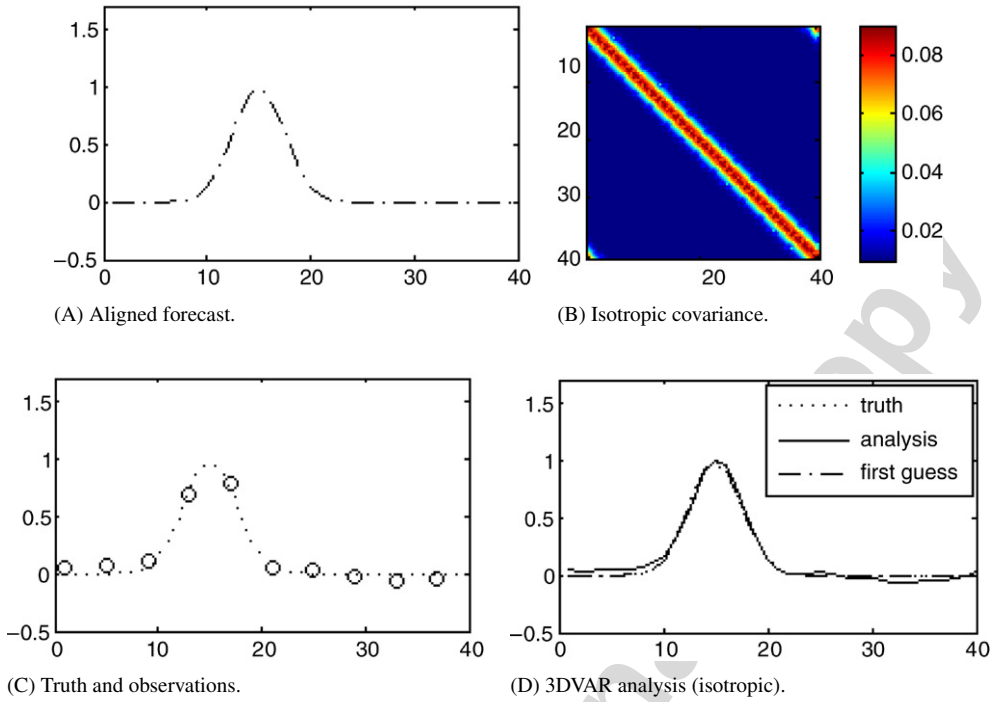
Fig. 10. Panel (A) shows a forecast and Panel (B) depicts the covariance computed from the forecast. Panel (C) shows the truth (dotted line) and observations. Panel (D) shows the aligned forecast (step 1), which becomes the first guess and Panel (E) shows the recomputed covariance. Panel (F) shows the 3DVAR solution (step 2).

Fig. 4(C)) to estimate the state, shown in panel D. Much like the case in Fig. 5 (right column), the performance is good when the position error is absent.

The two-step approach presents a significant computational saving over the one-step approach. The costs of computing the gradients with respect to positions in the one-step is comparable to one iteration of the alignment equation in the two-step. Similarly, the cost of computing the gradients with respect to amplitudes is comparable to the amplitude adjustment step in the two-step. As discussed, the one-step algorithm is more expensive from a dimensionality (memory) point of view. It also has a greater time complexity because the gradients with respect to amplitudes are computed at every iteration, whereas an equivalent computation is performed exactly once in the two-step. The BFGS scheme is capable of converging in fewer iterations than it takes the alignment step to converge. However, the time spent searching for the next amplitude-position adjustment together with the time spent in each iteration computing the gradient with respect to amplitudes makes it far more expensive than the two-step. The gradient computation with respect to displacements is $O(Nn \log n)$, for state size n and N ensemble members and the gradient computation with respect to amplitude is order $O(n^2 N)$. This order of magnitude difference cause makes the one-step algorithm slower in our experiments by an order of magnitude. For this reason and the fact that the two-step algorithm makes no use for the background error covariance in the displacement equation makes it feasible to use the alignment formulation as a *preprocessor* for an operational data assimilation system.

6. Two-dimensional examples

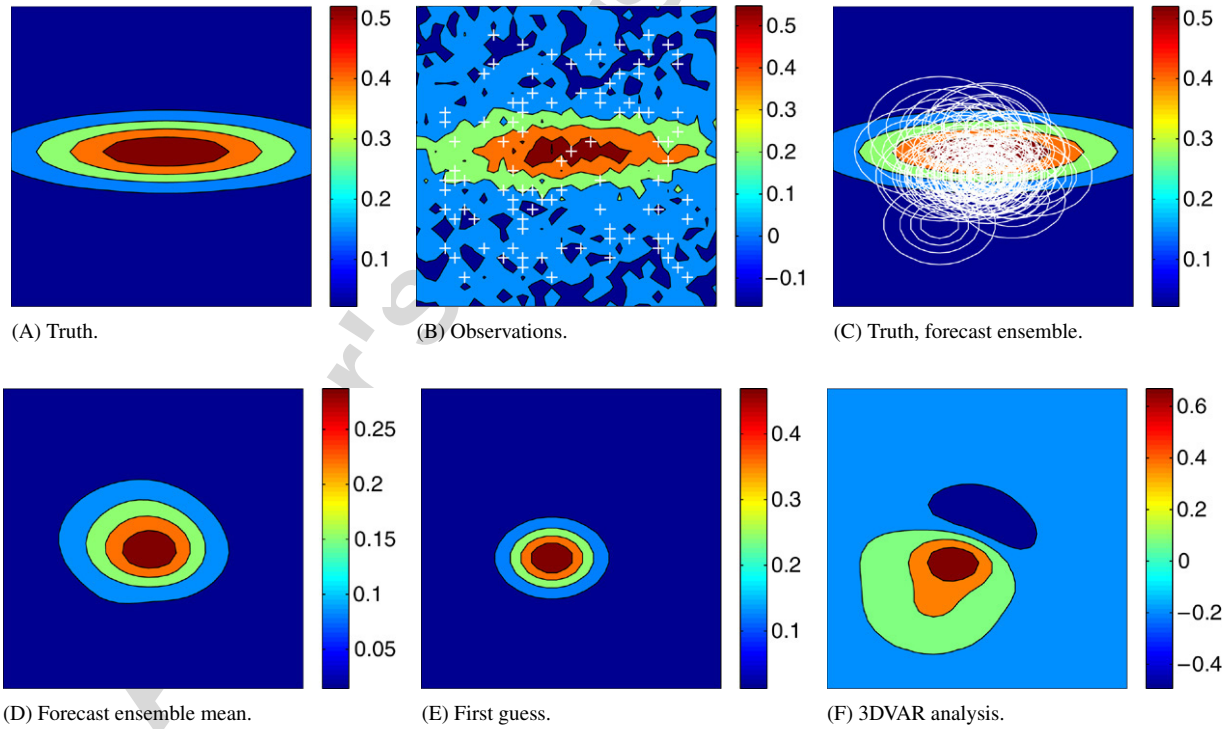
In this section, we examine the two-step approach in a two-dimensional problem. We use an ensemble of forecasts to construct the forecast uncertainty. These replicates have position and amplitude errors from truth. Comparisons will be made between 3DVAR and the two-step approach.



(A) Aligned forecast. (B) Isotropic covariance.

(C) Truth and observations. (D) 3DVAR analysis (isotropic).

Fig. 11. Panel (A) shows the aligned forecast (same as Fig. 10(D)) and Panel (B) depicts the associated isotropic background error covariance (same as Fig. 4(C)). Panel (C) shows the truth (dotted line) and observations (same as Fig. 10(C)). Panel (D) shows the 3DVAR solution with aligned first-guess and isotropic background error covariance.

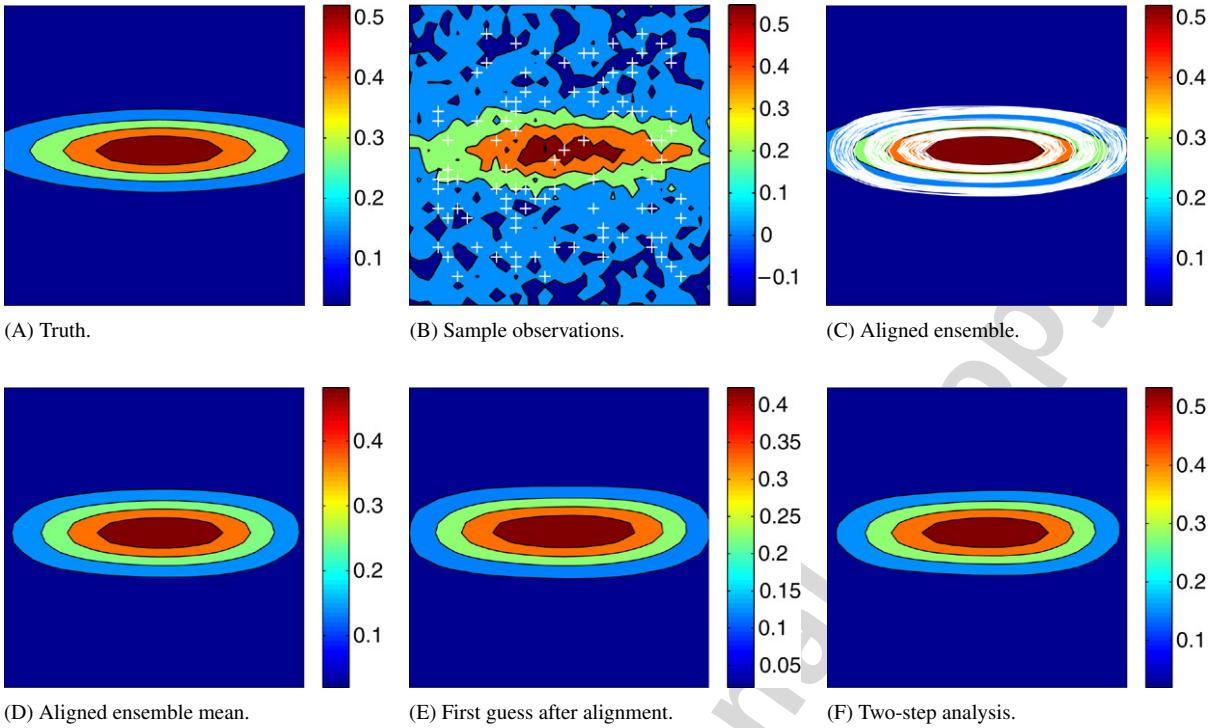


(A) Truth. (B) Observations. (C) Truth, forecast ensemble.

(D) Forecast ensemble mean. (E) First guess. (F) 3DVAR analysis.

Fig. 12. Two-dimensional example of data assimilation under position errors. Panel (A) is a true vortex field, observed at sparse locations (+ sign) shown in panel (B). Panel (C) shows an overlay of the forecast ensemble (white rings). Panel (D) depicts the forecast ensemble mean. The first guess is selected to be a coherent vortex of similar structure as the observations from the ensemble, as shown in panel (E). Panel (F) depicts the result of a deterministic assimilation scheme.

Fig. 12(A) depicts the vorticity contours of a true vortex field of size 32×32. Its peak amplitude is 0.6 non-dimensional units and it is elliptical in shape. Fig. 12(B) depicts observed locations (+), and the observation field is generated by adding to truth a 10% (of peak amplitude) uncorrelated noise field in space. Only 9.3% of the state is observed (96 observations), at



(A) Truth. (B) Sample observations. (C) Aligned ensemble.

(D) Aligned ensemble mean. (E) First guess after alignment. (F) Two-step analysis.

Fig. 13. This figure demonstrates a two-dimensional example of data assimilation by field alignment. Panels (A) and (B) are the same as Fig. 12. Panel (C) is the result of field alignment. This is reflected in panel (D) with better forecast ensemble mean. The first guess is selected from the ensemble as shown in panel (E). Panel (F) depicts the result of a deterministic assimilation scheme after alignment.

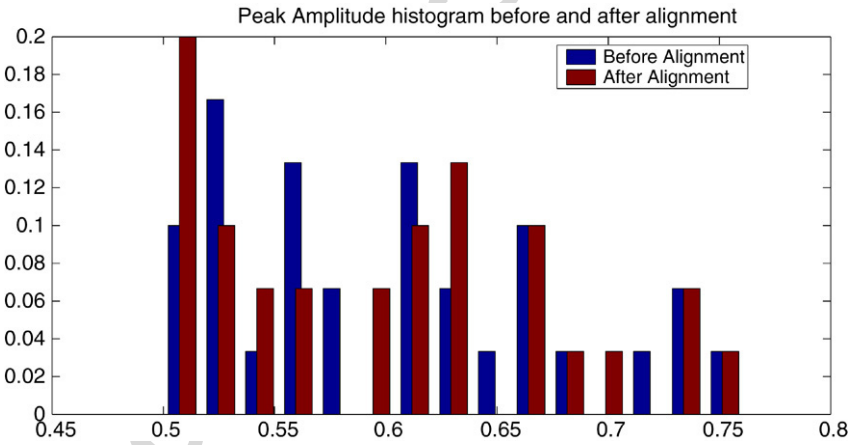


Fig. 14. The distribution of peak amplitudes in the forecast ensemble before and after the alignment step. Even though the ensemble members change shape (Fig. 13(C)) their amplitudes retain the spread that was in the pre-alignment ensemble.

randomly selected points. In particular, about 18 observations lie on significant contour levels of the vortex.

Fig. 12(C) depicts the contours (white) of a forecast ensemble (30 ensemble members) overlaid on truth. They are circular in shape and reasonably "cover" truth. All the ensemble members have position errors from truth as well as amplitude errors. Their peak amplitude can vary by as much as 25% and their position by 15 pixels. Fig. 12(D) depicts the forecast ensemble mean, which is broader than most forecast ensemble members and has a different shape than truth and its amplitude is far diminished from truth (by 50%). Fig. 12(E) depicts the first guess used for a 3DVAR procedure. The background error covariance (B) is computed from the forecast ensemble. Following the peak amplitude of 0.6 and 10% noise, the observation uncertainty is constructed from a 0.06 standard deviation, iid. Fig. 12(F) depicts the analysis produced by 3DVAR. The distortion of the vortex can be explained by the fact that position errors produce a poor representation of the background error covariance.

Fig. 13 depicts the same example using the proposed two-step approach. Fig. 13(A), (B) are identical to corresponding plots in Fig. 12. Fig. 13(C) depicts the *aligned* ensemble, using the alignment formulation developed in the previous section. In particular, Eq. (26) is used to align each ensemble member with a perturbed observation (of which Fig. 13(B) is one sample). As Fig. 13(C) shows, the forecast ensemble has both variations in

size and shape from truth and using the available observations has deformed its shape. Fig. 14 shows a histogram (distribution) of peak amplitude before and after alignment. It shows in particular that the amplitude spread is retained after alignment. This is useful to see because it shows that the ensemble does not over-fit the observations and retains its amplitude variability even as the forecast replicates have adjusted their shape to match the observations. The shape deformation is non-trivial in that it includes an expansion of the forecast replicates, which is a form of non-smooth deformation only possible because the smoothness and non-divergence are weak constraints.

The ensemble mean after alignment is shown in Fig. 13(D). An ensemble member is selected as the first guess for 3DVAR, as shown in Fig. 13(E). The analysis is shown in Fig. 13(F). The reason why the two-step approach works is because the aligned ensemble may be considered to be a far better depiction of background uncertainty than the one containing position errors. Experiments using the EnKF to produce the analysis leaves us with the same conclusions, as the one-dimensional examples also show. These are not repeated here.

We examine in an experiment how much the two-step improves over the 3DVAR solution. In this experiment, we generate a true vortex (similar to the forecasts shown in Fig. 12). This vortex is a two-dimensional (unnormalized) Gaussian. A spatially uncorrelated noise of standard deviation 0.01 is added to it, and measured at sparse randomly s elected locations, covering 6.8% of the domain. A first guess is generated with a vortex center randomly positioned in the domain. An ensemble of forecasts is generated around the first guess by perturbing the Gaussian standard deviation (0.2) and position (5 pixels). An analysis is produced by computing the forecast error covariance from the ensemble using 3DVAR (PSAS variation, see Example 1). The two-step approach is applied to the ensemble. The aligned ensemble is used to compute a new background error covariance, and the aligned first-guess is used in 3DVAR. Analysis errors in both cases are computed from truth using the root mean squared error (RMS). Fig. 15 shows 100 such simulations. The X-axis is the analysis error without alignment and the Y-axis is the analysis error using the two-step approach. This scatter-plot therefore indicates that the two-step produces consistently better analyses.

6.1. Implementation

Gradients in the forcing terms of Eq. (26) are computed using central differences. Laplacians in the displacement pde are implemented using biharmonic operator, or the biLaplacian. The displacement pde is solved using a spectral method. Displacement updates to the state in its evolution to $\hat{X}^f$ are done using bicubic interpolation.

The nominal boundary condition for the displacement pde is homogeneous, $q_i = 0, i \in \text{Boundary}(\Omega)$. In general, the boundary condition depends on the problem being solved. For example if a doubly periodic domain is assumed, then the displacement pde must have circular conditions. Typical experiments were carried out on domains of size 32×32 with extended boundaries. A 32×32 problem is embedded in a larger

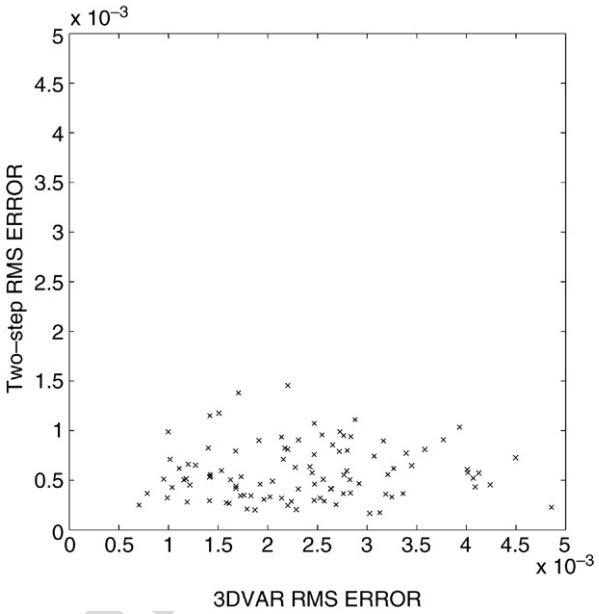


Fig. 15. A comparison of the analysis errors between the two-step approach and deterministic shows substantial improvements. The X-axis shows the analysis (RMS) error for 3DVAR and the Y-axis shows the corresponding error for the two-step approach.

computational grid, that is the boundaries are considered "far away".

To demonstrate the sensitivity to parameters w_1 and w_2, we construct an example shown in Fig. 16. This example contains a field with two vortices. The left pane is the forecast, while the middle pane is truth shown with station observations. The peak amplitude of the field is 1 non-dimensional unit and the smallest amplitude is zero. The domain is 32×32 and every fourth pixel is observed without noise.

In the *sequential approach*, w_1 and w_2 only influence the alignment solution and do not directly influence amplitude adjustments. This is in contrast to their role in the one-step algorithm, where the weights control the relative influence of amplitude and position errors. Here, we choose to set $w_2 = \frac{w_1}{3}$, for in this case there is an analogy to interpreting w_1 as a viscosity parameter. Although w_2 can be changed as well, but when $w_1 = w_2$, the constraints lead to a singularity in the displacement equation and must be avoided. For most practical implementations, $w_2 = \frac{w_1}{3}$ suffices. To pick a value of w_1 we conducted simulations that could displace a tight Gaussian of unit peak amplitude and standard deviation equal to 0.8 by 10 pixels in three iterations or more. This was satisfied by a value of $w_1 \geq 10^{-4}$. We then used this value as a lower bound together with $w_2 = \frac{w_1}{3}$.

To examine the sensitivity to the parameters w_1 and w_2, we examine the amplitude error (rms) between the aligned and true state, as well as the number of iterations the alignment algorithm takes to converge. Fig. 17 shows the error at convergence, that is the error between the aligned first-guess field and truth, when observations are sparse, but noise-free. Error is defined as the absolute value of the maximum error observed between the aligned first-guess field and truth. Convergence is defined as the maximum error between the

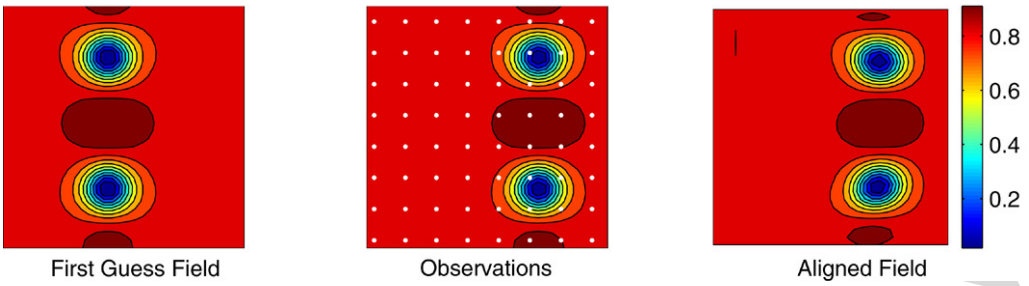


Fig. 16. To examine the sensitivity to weights w_1 and w_2, we use a two-vortex example. The panel on the left is the forecast field, the panel in the middle depicts truth, which is observed without noise at station observations (dots). The panel on the right depicts the aligned forecast.

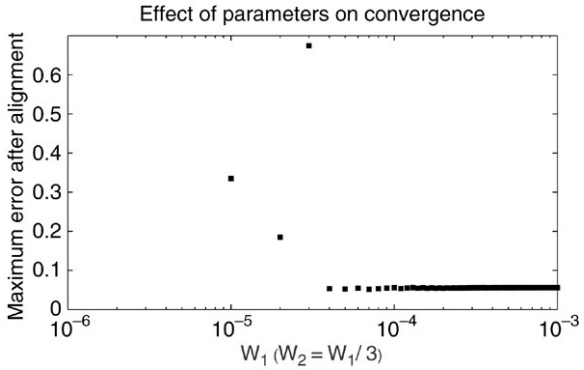


Fig. 17. This graph depicts the error between aligned first-guess and true fields as a function of the value of w_1, when $w_2 = \frac{w_1}{3}$. It demonstrates that the alignment solution is robust to moderate changes in w_1. The x-axis is on a log-scale.

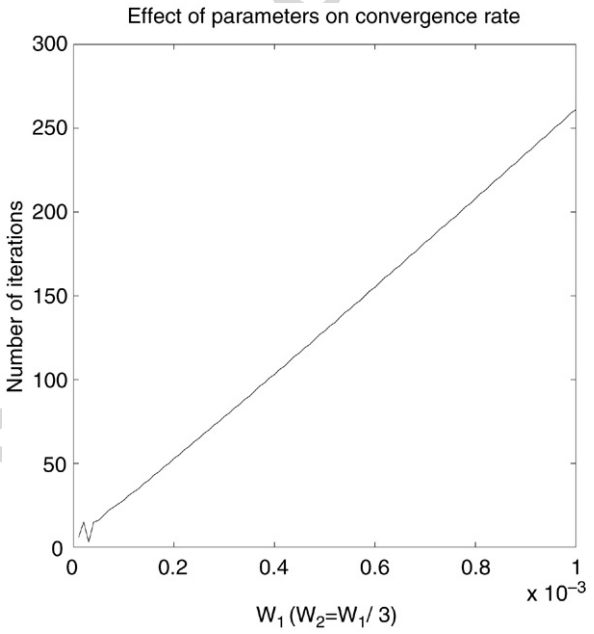


Fig. 18. This graph depicts the number of iterations (y-axis) it takes to align the first guess with the observed field as a function of the value of w_1 (x-axis) when $w_2 = \frac{w_1}{3}$.

two fields reaching 0.05, or if 1000 iterations have elapsed. Fig. 17 shows the error convergence for w_1 varying from 10^{-5} to 10^{-3} (100 values total). This graph shows that the first three parameter settings (10^{-5}, 2×10^{-5} and 3×10^{-5} respectively) are insufficient to constrain the solution; in fact it blows up. This is indicated by the large residuals and the field is not aligned. Thereafter, we see that the field aligns itself, to very nearly the same state, up to the convergence criteria. As w_1 is increased, the weights associated with the constraints go up. Since the only adjustments they affect are displacements, larger weights can make the alignment more cohesive, forcing homogeneous motions at larger scales. However, in our example, the embedding of the domain in a larger computation grid also achieves this effect by forcing the displacement field to smaller wave numbers in the spectral solution.

Therefore, the predominant effect of increasing weights is that the instantaneous displacement at every iteration gets smaller and smaller, thus taking longer for convergence. This can be seen in Fig. 18. As w_1 increases so does the time to alignment. The solution depicted in Fig. 16 (right panel) and Fig. 13 was produced with a value of $w_1 = 10^{-4}$. Thus, the alignment solution is robust to moderate changes in the parameters and not very sensitive to them.

7. Discussion and conclusions

Position errors are ubiquitous in forecasting, and affect verification. The sources of position errors are poorly understood and reducing them to component sources is difficult. But do they matter? We think so because the presence of position errors can violate the assumptions driving current assimilation techniques. When they do, the analysis will be unsurprisingly bad. We believe that attention must be paid to this problem and the proposed formulation demonstrates how position information can be incorporated into an analysis procedure.

Whilst the one-step algorithm, synthesized directly from the Bayesian formulation, is interesting, it is not scalable and hence subject to the criticism that this method may add yet another layer to already expensive assimilation implementations for large-scale problems. The one-step algorithm also required us to construct statistical error covariances that may not be feasible in reality.

The composite objective can be solved in two steps as an approximation. At least in one-dimensional examples, we see little difference between the one-step and two-step solutions. This makes this algorithm attractive from a practical point of view. We think that when the analyst decides that position

errors are significant, as a quality control scheme may suggest, applying the alignment procedure would be useful to mitigate the effects of a poor background. This is feasible because once we partition data assimilation in two steps, alignment is essentially a preprocessing procedure to classical data assimilation. Thus, forecast fields, after alignment, can be used to construct new background error statistics (post hoc). This makes what we propose useful with or without ensemble forecasts.

We have also argued that the proposed alignment method offers advantages over feature-based methods. Feature-based methods detect key structures: in the case of vortices this may be the vortex center. There are three main issues that contrast the current approach with a feature-based approach. First, it is not clear that features are well defined in model fields. In observations, although features could be detected from satellite imagery, it is not clear that this can be done without some preprocessing of sparse point observations as we have used. Essentially this requires the user to construct a full field from sparse observations. Second, even if features can be detected, it is unclear how to align the rest of the field. Aligning features is not the same as aligning fields because features are sparse by definition. The number of vortex or pressure centers, significant contour levels, are far fewer than the number of nodes in the state and the field deformation may not be a simple translation of features. Third, our approach can, in principle, be extended to incorporate detected features. This can be done by providing boundary conditions to the alignment Eq. (26) by simply setting the displacements to known values for certain grid nodes and using the alignment procedure to fill-in displacement values at remaining nodes. We have not demonstrated this but will do so in future work.

We have also examined extensions of the two-step algorithm to multivariate fields and are developing extensions to 3D fields. In the former case, using pressure and velocity fields, we see that alignment does not perturb a preexisting physical balance. These extensions will be reported in subsequent work demonstrating the utility of the two-step on large-scale problems and in a dynamic setting.

The analysis produced with the use of alignment is demonstrably much more realistic than a 3DVAR or EnKF solution. In essence, the power of this algorithm lies in the very simple and robust constraints that can be used to ameliorate the position error problem affecting assimilation. The sequential framework is a good approximation, but more efficient approaches to solve for displacement–amplitude adjustments jointly may prove attractive.

Acknowledgments

This material is based on work supported in part by an NSF ITR grant, No. 0121182. The authors thank Dr. Ragoth Soundararajan for comments, and Prof. John Marshall for his enthusiastic support and encouragement of this research.

References

[1] G.D. Alexander, J.A. Weinman, J.L. Schols, The use of digital warping of microwave integrated water vapor imagery to improve forecasts of marine extratropical cyclones, Mon. Weather Rev. 126 (1998) 1469–1495.
[2] J. Dudhia, A non-hydrostatic version of the penn state/ncar mesoscale model: Validation tests and simulation of an atlantic cyclone and cold front, Mon. Weather Rev. 121 (1993) 1493–1513.
[3] H.J. Thiebaux, P.R. Julian, G.J. DiMego, Areal versus collocation data quality control, in: Intl. Symp. on Assimilation of Observations in Meteorology and Oceanography, Clermont-Ferrand, France, WMO, 1990, pp. 255–260.
[4] C.D. Jones, B. MacPherson, A latent heat nudging scheme for assimilation of precipitation data into an operation mesoscale model, Meteorol. Appl. (1997) 269–277.
[5] R.N. Hoffman, Z. Liu, J. Louis, C. Grassotti, Distortion representation of forecast errors, Mon. Weather Rev. 123 (1995) 2758–2770.
[6] R.N. Hoffman, C. Grassotti, A technique for assimilating ssm/i observations of marine atmospheric storms: Tests with ecmwf analyses, J. Appl. Meteorol. 35 (1996) 1177–1188.
[7] A.J. Mariano, Contour analysis: A new approach for melding geophysical fluids, J. Ocean. Atmos. Technol. 7 (1990) 285–295.
[8] K. Puri, G. Holland, Numerical track prediction models, in: G. Holland (Ed.), Global Guide to Tropical Cyclone Forecasting, Bureau of Meteorology Research Center, 2000. http://www.bom.gov.au/pubs/tcguide/globa_guide/intro.htm.
[9] K.A. Brewster, Phase-correcting data assimilation and application to storm-scale numerical weather prediction. Part i: Method description and simulation testing, Mon. Weather Rev. 131 (2003) 480–492.
[10] R.P. Dee, A.M. da Silva, Data assimilation in the presence of forecast bias, Q. J. Roy. Meteorol. Soc. 124 (545) (1998) 269–296.
[11] C. Wunsch, 2003 (personal communication).
[12] A.H. Jazwinski, Stochastic Processes and Filtering Theory, Academic Press, 1970.
[13] S.E. Cohn, A.D. Silva, J. Guo, M. Sienkiewicz, D. Lamich, Assessing the effects of data selection with the dao physical-space statistical analysis system, Mon. Weather Rev. 126 (1998) 2913–2926.
[14] P. Courtier, Variational methods, J. Meteorol. Soc. Jpn. 75, 211–218.
[15] A.C. Lorenc, Analysis method for numerical weather prediction, Q. J. Roy. Meteorol. Soc. 112 (1986) 1177–1194.
[16] E. Kalnay, Atmospheric modeling, in: Data Assimilation and Predictability, Cambridge University Press, 2003.
[17] G. Evensen, The ensemble Kalman filter: Theoretical formulation and practical implementation, Ocean Dynam. 53 (2003) 342–367.
[18] A. Tikhonov, V.Y. Arsenin, Solutions of Ill-Posed Problems, Wiley, New York, 1977.
[19] T. Nehrkorn, R.N. Hoffman, C. Grassotti, J.F. Louis, Feature calibration and alignment to represent model forecast errors: Empirical regularization, Q. J. Roy. Meteorol. Soc. 129 (2003) 195–218.
[20] C.G. Broyden, The convergence of a class of double-rank minimization algorithms, J. Inst. Math. Appl. 6 (1970) 76–90.
[21] R. Fletcher, A new approach to variable metric algorithms, Comput. J. 13 (1970) 317–322.
[22] D. Goldfarb, A family of variable metric updates derived by variational means, Math. Comput. 24 (1970) 647–656.
[23] D.F. Shanno, Conditioning of quasi-Newton methods for function minimization, Math. Comput. 24 (111) (1970) 647–656.
[24] G. Wabha, J. Wendelberger, Some new mathematical methods for variational objective analysis using splines and cross-validation, Mon. Weather Rev. 108 (1980) 1122–1143.
[25] R. Daley, Atmospheric Data Analysis, Cambridge University Press, 1994.