EARLY ONLINE RELEASE

This is a PDF of a manuscript that has been peer-reviewed and accepted for publication. As the article has not yet been formatted, copy edited or proofread, the final published version may be different from the early online release.

This pre-publication manuscript may be downloaded, distributed and used under the provisions of the Creative Commons Attribution 4.0 International (CC BY 4.0) license. It may be cited using the DOI below.

The DOI for this manuscript is
DOI:10.2151/jmsj.2019-067

J-STAGE Advance published date: September 24th, 2019

The final manuscript after publication will replace the preliminary version at the above DOI once it is available.
Ensemble Kalman Filtering Based on Potential Vorticity for Atmospheric Multi-scale Data Assimilation

Tadashi TSUYUKI

Meteorological College
Japan Meteorological Agency, Kashiwa, Japan

1) Corresponding author: Tadashi Tsuyuki, Meteorological College, 7-4-81 Asahi-cho, Kashiwa, 277-0852, JAPAN.
Email: ttsuyuki@mc-jma.go.jp
Tel: +81-4-7144-7185
A multi-scale data assimilation method for the ensemble Kalman filter (EnKF) is proposed for atmospheric models in cases with insufficient observations of fast variables. This method is based on the conservation and invertibility of potential vorticity (PV). The dynamical state variables in the free atmosphere of forecast ensemble members are decomposed into balanced and unbalanced parts by applying PV inversion to the PV anomalies computed from spatially smoothed state variables. The mass variables of the two parts are adjusted to remove additional sampling errors introduced by this decomposition. The forecast error covariances between those parts are ignored in the Kalman gain to suppress spurious error correlations. This approximation makes it possible to apply different covariance localizations to each part. The Kalman gain thus obtained is used to assimilate observations.

The performance of the proposed method is demonstrated with a shallow water model through twin experiments in a perfect model scenario. Results using the same localization radius for the two parts show that the proposed EnKF is superior in the accuracy of the analysis to a conventional EnKF unless the ensemble size is sufficiently large. It is found that the adjustment of mass variables is necessary to outperform the conventional EnKF. The benefits of the PV inversion using the Bolin-Charney balance over the quasi-geostrophic inversion are marginal in the experiments.
Keywords  ensemble Kalman filter; multi-scale data assimilation; potential vorticity
1. Introduction

With rapidly increasing computational power, current advanced numerical models of the global atmosphere have higher spatial resolutions. Among them are a cloud-resolving global nonhydrostatic model, NICAM (Nonhydrostatic Icosahedral Atmospheric Model; Satoh et al., 2008, 2014), and a gravity wave-resolving middle atmosphere model, JAGUAR (Japanese Atmospheric General Circulation Model for Upper Atmosphere Research; Watanabe et al., 2008; Watanabe and Miyahara, 2009). Those multi-scale global atmospheric models exhibit a wide range of dynamic and thermodynamic phenomena ranging from planetary scales up to cumulus convection or gravity waves, and they have been used to study scale interactions and predictability. The explicit representation of small scales in space and time in high-resolution atmospheric models should improve data assimilation and numerical weather prediction (NWP), although those scales are usually not fully observed. Generally, small scales change rapidly and large scales evolve slowly, so that small (large) scales and fast (slow) variables will be interchangeably used in the following unless otherwise stated.

A conventional approach to deal with multi-scales in data assimilation in meteorology is nesting of numerical models: low frequent observations of large scales, such as radiosonde and polar-orbiting satellite data, are assimilated into a large-area model with a low resolution, and highly frequent observations of small scales, such as Doppler radar data, are assimilated into a small-area model with a high resolution, of which lateral boundary
conditions are provided by the large-area model. Though this approach has been very successful in the operational NWP, there are a couple of problems in data assimilation. One of them is associated with data selection. Observations usually have the information of both large and small scales. If those data are assimilated into only one of the models, part of the information is discarded. On the other hand, it is difficult for the small-area model to use all information of observational data, because observations outside the domain of this model are not assimilated. The other problem is the absence of feedback of small-scale observations to the large-area model unless two-way nesting is adopted.

Multi-scale data assimilation, in which observations of small and large scales are simultaneously assimilated into a large-area model with a high resolution, is free from those problems. A problem with multi-scale data assimilation is that high-resolution observations of small scales are usually available only in limited regions, as mentioned above. As a result, the background state of small scales is inaccurate in most of the domain of the large-area model. Even if those data are available over the globe with a high temporal resolution, frequent assimilation of the data into the large-area model may be impractical. As atmospheric phenomena of smaller scales have increasingly limited predictability, the background state of small scales is increasingly inaccurate if assimilation is less frequent.

Lorenc and Payne (2007) discuss those issues in the context of four-dimensional variational data assimilation (4D-Var). They introduce the concept of statistical 4D-Var to cope with a wide range of scales, in which a mean forecast equation is used instead of a
deterministic forecast equation. However, an ensemble-based approach would be more promising to address predictability issues in data assimilation.

Ballabrera-Poy et al. (2009) investigate the accuracy of an ensemble Kalman filter (EnKF) when observations of fast and slow variables are simultaneously assimilated into the Lorenz-96 two-scale model (Lorenz, 1996). Their results demonstrate that assimilation of a few observations of fast variables significantly degrades the filter performance due to spurious error correlations. Suppose that the amplitude of a fast variable depends on a slow variable, such as in interactions between cumulus convection and its environment, or that a small-scale variable is advected by large-scale flow, such as in interactions between waves and mean flow. In addition, the background state of the fast variable is supposed to have large errors due to an insufficient number of observations, so that the amplitude and phase of the fast variable are dispersed among forecast ensemble members. As a result, even if the fast variable strongly depends on the slow variable, forecast error correlations between the fast and slow variables should be very weak. However, if the ensemble size of an EnKF is not sufficiently large, then spurious error correlations between them are likely to be computed, and assimilation of an insufficient number of observations of the fast variable may degrade the analysis of the slow variable.

We can avoid this problem by ignoring computed forecast error correlations between fast and slow variables, since these correlations are expected to be very weak in cases with insufficient observations of fast variables. To ignore the correlations, a procedure is needed
for extracting slow variables from state variables of the atmosphere. We apply potential vorticity (PV) inversion (Hoskins et al., 1985) for this purpose. In adiabatic and frictionless flows, PV evolves according to advection. Therefore, if the balanced flow derived by PV inversion is of large scales, it changes slowly with time, and the residual flow is of small scales or evolves fast. The latter flow represents the unbalanced flow. The neglect of forecast error correlations between the balanced and unbalanced flows will alleviate the problem demonstrated by Ballabrera-Poy et al. (2009). The purpose of the present study is to formulate this multi-scale data assimilation method for the EnKF and to demonstrate its performance with a shallow water model through twin experiments in a perfect model scenario. The method is hereafter referred to as the PV-based EnKF. Because decomposing state variables into two parts may cause additional sampling errors in an EnKF, some adjustment of decomposed variables is necessary. It will be shown that adjustment of mass variables is needed for the PV-based EnKF to outperform a conventional EnKF.

Application of PV to data assimilation has been investigated for variational data assimilation by several authors (Cullen, 2003; Wlasak et al., 2006; Bannister and Cullen, 2007; Katz et al., 2011). In the operational variational data assimilation, the background error covariance matrix is huge, and many types of approximations must be introduced. One of them is the parameter transform (Bannister, 2008), in which model variables are transformed into control variables whose errors are assumed to be uncorrelated. A natural
way is to partition the flow into balanced and unbalanced parts. Vorticity or a stream function has been chosen as the balanced variable in operational 4D-Var systems. According to the geostrophic adjustment theory, however, it is on horizontal scales shorter than the Rossby deformation radius that rotational flow can be regarded as balanced. This approximation breaks down on scales larger than the Rossby deformation radius, and a mass variable becomes suitable for the balanced variable. The above authors propose PV-based transformations and show their benefits including successful production of a set of uncorrelated control variables. Cullen (2003) and Bannister and Cullen (2007), however, find a difficulty in implementing PV-based transformations to operational variational data assimilation systems, because the three-dimensional elliptic equations of linear PV inversion are ill-conditioned. This difficulty remains an issue to be solved for the operational implementation of PV-based control variables. Thus, the concept of the PV-based EnKF is different from that of the above method for variational data assimilation, although both methods make use of PV for data assimilation.

The remainder of this paper is organized as follows. Section 2 introduces the formulation of the multi-scale data assimilation method for the EnKF. Section 3 describes the design of data assimilation experiments with a shallow water model. Results of the experiments are presented in Section 4, and a summary and a couple of issues are mentioned in Section 5.

2. Method
The proposed multi-scale data assimilation method for the EnKF is based on the conservation and invertibility of PV, and consists of three components: (i) decomposition of the dynamical state variables in the free atmosphere of forecast ensemble members into balanced and unbalanced parts by applying PV inversion to the PV anomalies computed from spatially smoothed state variables, (ii) adjustment of the mass variables of the two parts to remove additional sampling errors introduced by this decomposition, and (iii) approximation to the Kalman gain by ignoring forecast error covariances between the two parts. The Kalman gain thus obtained is used to assimilate observations.

2.1 PV inversion of large scales

PV is a functional of the dynamical state variables $x$ of the atmosphere, and it is written as $P[x]$. In PV inversion, the balanced part of $x$, which is denoted by $x_b$, is derived by solving the equation

$$P[x_b] = P[x] \quad (1)$$

with the equations of balance conditions under appropriate boundary conditions. In adiabatic and frictionless flows, PV is conserved on fluid particles and evolves according to advection. Assume that $x_b$ is uniquely determined and is a smooth functional of PV. Then, if the spatial scale of PV is large, $x_b$ changes slowly with time. Therefore, $x_b$ is a suitable candidate for slow variables. The unbalanced part $x_u$, which is defined as

$$x_u = x - x_b \quad (2)$$
evolves without the constraint of PV conservation. If the PV of $x_u$ is defined as

$$P[x_b + x_u] - P[x_b]$$

$x_u$ has no PV. Thus, the physical mechanisms of the evolution of $x_b$ and $x_u$ are fundamentally different, and they evolve independently in linear approximation.

Strong diabatic heating is associated with latent heating of water, which is often localized in small regions of intense precipitation. Friction is almost concentrated in the planetary boundary layer except for the vicinity of jet streams, fronts, and convective clouds. Consequently, PV is nearly conserved in the free atmosphere, and the above partition of the dynamical state variables should be applied in the free atmosphere only. The proposed data assimilation method is therefore not suitable for estimating state variables in the planetary boundary layer.

Since PV evolves according to advection, a balanced flow of larger scale evolves more slowly. Therefore, we apply PV inversion to large-scale PV to extract slow variables from the dynamical state variables of the atmosphere. To obtain large-scale PV, spatial smoothing is needed. There are two procedures for this purpose. The first one is to compute PV from state variables and then to smooth the PV. The other one is to smooth state variables and then to compute PV from the smoothed state variables. Since PV is a nonlinear functional of state variables, these procedures lead to fundamentally different definitions of large-scale PV. Keyser and Rotunno (1990) discuss this issue in detail and recommend the latter procedure for a practical reason; a grid point dataset of fluid is not free from smoothing. They also point out that the idea behind the former procedure is that
PV behaves like a tracer. However, PV is not a passive tracer but an active tracer. The impermeability theorem for PV (Haynes and McIntyre, 1987, 1990) illustrates that there is a fundamental difference between the behavior of PV and that of chemical tracers. The above discussion suggests that the latter procedure may be preferable, and we take this option. The introduction of spatial smoothing implies that a low-resolution model can be used for the inversion of large-scale PV. The use of a low-resolution model reduces the computational cost of PV inversion and alleviates the difficulty in solving the ill-conditioned three-dimensional elliptic equations. Note that if the large-scale PV is used for deriving the balanced part, the unbalanced part may contain a part of the state variables that evolve according to PV conservation.

In PV inversion for the three-dimensional atmosphere, hydrostatic equilibrium is usually assumed in the vertical, and the Ertel PV with the hydrostatic approximation is inverted by using balance conditions such as the Bolin-Charney balance (Bolin, 1956; Charney, 1955, 1962). This balance condition is written in pressure coordinates as

\[ \nabla^2 \phi_b = \nabla \cdot f \nabla \psi_b + 2 \left[ \frac{\partial^2 \psi_b}{\partial x^2} \frac{\partial^2 \psi_b}{\partial y^2} - \left( \frac{\partial^2 \psi_b}{\partial x \partial y} \right)^2 \right] \]

(3)

\[ \mathbf{v}_b = \left( -\frac{\partial \psi_b}{\partial y}, \frac{\partial \psi_b}{\partial x} \right), \]

(4)

where \( \phi_b \) is the balanced part of geopotential departure from its reference state, \( \mathbf{v}_b \) is the balanced part of horizontal velocity, \( \psi_b \) is its stream function, \( f \) is the Coriolis parameter, \( \nabla \) is the two-dimensional nabla operator, and \( \nabla^2 \) is the two-dimensional Laplacian.
operator. There is no balanced part of divergence. This balance is a generalization of the
gradient wind equilibrium. The uniqueness of solutions and the mass conservation in PV
inversion will be discussed in Subsection 3.2, specifically for the shallow water equations
with the periodic boundary condition.

Several methods of PV inversion have been proposed. Davis and Emanuel (1991)
present an algorithm of PV inversion with the Bolin-Charney balance in pressure
coordinates on the globe. In their algorithm, a set of nonlinear elliptic partial differential
equations are solved iteratively. Although it has been widely used for analyzing synoptic
disturbances in middle latitudes (e.g., Stoelinga, 1996; Korner and Martin, 2000), the
three-dimensional elliptic equations are ill-conditioned, and there is some degree of
arbitrariness in setting boundary conditions. McIntyre and Norton (2000) formulate two
hierarchies of PV inversion for the shallow water equations: direct inversions and normal
mode inversions. The first-order direct inversion corresponds to the algorithm of Davis and
Emanuel (1991). They demonstrate that normal mode inversions are more accurate than
direct inversions. Arbogast et al. (2008) propose another method of PV inversion, which
makes use of digital filter initialization. Their method searches the minimum of a quadratic
cost function that measures the difference from initialized flows subject to the constraint
that PV does not change. The digital filter initialization scheme of Lynch et al. (1997) is
iteratively performed during the search of the minimum, and boundary conditions are
required only for a Lagrangian multiplier function. PV inversion methods based on
initialization avoid solving the ill-conditioned three-dimensional elliptic equations for balanced variables. On the other hand, Viudez and Dritschel (2004) introduce the concept of optimal PV balance, where the balanced flow is a solution of dynamical equations and is minimally free of the unbalanced flow. Their PV inversion method is based on a Lagrangian description of fluid dynamics, and the balanced flow is obtained iteratively in a cycle of backward and forward integrations of a numerical model, during which PV is restored in every iteration loop.

In the present study, data assimilation experiments are conducted using the shallow water equations, and the PV for the equations is used instead of the Ertel PV. We choose the first-order direct inversion of McIntyre and Norton (2000), but the quasi-geostrophic inversion is also employed for comparison. In the latter inversion, geostrophic balance is adopted as the balance condition, and the quasi-geostrophic PV for the shallow water equations is inverted. The quasi-geostrophic PV is conserved on virtual fluid particles moving at the velocity of geostrophic wind. Therefore, if geostrophic wind is not very different from actual wind, the balanced flow derived from the quasi-geostrophic PV also evolves slowly. It implies that we can use the quasi-geostrophic inversion to derive the balanced part of the dynamical state variables.

2.2 Adjustment of mass variables

The decomposition of the dynamical state variables by PV inversion may introduce
additional sampling errors. Let us take the Bolin-Charney balance in pressure coordinates of Eq. (3) as an example. The geopotential departure $\phi$ is called the mass variable in this subsection, and let the overbar denote the horizontal area average of a state variable. As the area average of the balanced mass variable $\overline{\phi_b}$ does not appear in Eq. (3), it is expected that there are no forecast error correlations between $\overline{\phi_b}$ and $\nu_b$. On the other hand, as they are derived from the same PV anomaly, the existence of weak error correlations cannot be ruled out. In the present study, we assume that possible error correlations are so weak that the forecast covariances between them can be ignored:

$$\langle \Delta \overline{\phi_b} \Delta \nu_b \rangle \approx 0,$$  \hspace{1cm} (5)

where $\Delta A$ denotes the forecast error in $A$, and a bracket represents the expectation operation. Results presented in Subsection 4.2 will indirectly justify this assumption.

As sampling errors are inevitable, Equation (5) generally does not hold in an EnKF. Then assimilation of velocity observations produces an analysis increment of $\overline{\phi_b}$, and it may lead to a degraded analysis of the mass variable. Equation (5) also suggests that the error correlation between $\overline{\phi_b}$ and $\phi_b' = \phi_b - \overline{\phi_b}$ is weak, because $\phi_b'$ is determined by $\nu_b$ through Eq. (3). Large sampling errors in $\langle \Delta \overline{\phi_b} \Delta \phi_b' \rangle$ may also degrade the analysis when observations of the mass variable are assimilated.

To remove those sampling errors introduced by the decomposition, we adjust the balanced and unbalanced parts of the mass variable such that $\langle \Delta \overline{\phi_b} \Delta \nu_b \rangle$ and $\langle \Delta \overline{\phi_b} \Delta \phi_b' \rangle$ vanish. In the present study, the following modified balanced and unbalanced mass
variables, $\tilde{\phi}_b$ and $\tilde{\phi}_u$, are introduced:

\begin{align}
\tilde{\phi}_b &= \phi_b', \tag{6} \\
\tilde{\phi}_u &= \phi_u + \tilde{\phi}_b. \tag{7}
\end{align}

That is, the area-averaged mass is considered as an unbalanced variable, and it is transferred from the balanced part to the unbalanced one. Since $\tilde{\phi}_b$ always vanishes, Equation (5) exactly holds for $\tilde{\phi}_b$. Note that the total mass variable remains unchanged.

This procedure is hereafter referred to as the mass adjustment, and the new variables, $\tilde{\phi}_b$ and $\tilde{\phi}_u$, are used in place of the original variables, $\phi_b$ and $\phi_u$, in the following. Although the new set of balanced variables thus obtained does not exactly satisfy Eq. (1), its influence is negligible if $\tilde{\phi}_b \approx 0$.

\subsection{2.3 Approximation to Kalman gain}

In the case of linear dynamics, we can safely ignore forecast error correlations between the balanced and unbalanced parts, assuming that the influence of possible analysis error correlations between them almost vanishes in the presence of forcing and dissipation during time integration from the analysis. In the case of nonlinear dynamics, there may be forecast error correlations between the two parts through nonlinear interactions. However, if observations of the unbalanced part are insufficient, these correlations are expected to be very weak as described in the introduction. Then if the ensemble size of an EnKF is not sufficiently large, we can alleviate the degradation of analysis due to spurious error
correlations between the balanced and unbalanced parts by ignoring them in the Kalman gain.

By applying PV inversion and the mass adjustment to the forecast $\mathbf{x}_f$ of each ensemble member, the forecast error $\Delta \mathbf{x}_f$ of each member is decomposed as $\Delta \mathbf{x}_f = \Delta \mathbf{x}_b + \Delta \mathbf{x}_u$, where $\Delta \mathbf{x}_b$ and $\Delta \mathbf{x}_u$ are the forecast errors of the balanced and unbalanced parts, respectively, and the former satisfies Eq. (5). If the forecast error covariances between $\Delta \mathbf{x}_b$ and $\Delta \mathbf{x}_u$ are neglected, the forecast covariance matrix $\mathbf{P}^f = \left[ \left( \Delta \mathbf{x}_b \right)^T \left( \Delta \mathbf{x}_b \right) \right]$ is approximated by

$$\mathbf{P}^f \approx \mathbf{P}_b^f + \mathbf{P}_u^f$$

where

$$\mathbf{P}_b^f = \left[ \left( \Delta \mathbf{x}_b \right)^T \left( \Delta \mathbf{x}_b \right) \right]$$

(9)

$$\mathbf{P}_u^f = \left[ \left( \Delta \mathbf{x}_u \right)^T \left( \Delta \mathbf{x}_u \right) \right]$$

(10)

In an EnKF, the expectation operation is replaced by the unbiased estimate based on an ensemble. Then the Kalman gain $\mathbf{K}$ can be approximated by the sum of balanced and unbalanced components:

$$\mathbf{K} \approx \mathbf{K}_b + \mathbf{K}_u$$

(11)

where

$$\mathbf{K}_b = \mathbf{P}_b^f \mathbf{H}^T \left( \mathbf{R} + \mathbf{H} \mathbf{P}^f \mathbf{H}^T \right)^{-1}$$

(12)

$$\mathbf{K}_u = \mathbf{P}_u^f \mathbf{H}^T \left( \mathbf{R} + \mathbf{H} \mathbf{P}^f \mathbf{H}^T \right)^{-1}$$

(13)

$\mathbf{R}$ is the observation error covariance matrix, and $\mathbf{H}$ is the linearized observation operator.
Li et al. (2015) propose a similar decomposition of the background error for multi-scale variational data assimilation, in which state variables are decomposed into two spatially distinct scales and the errors of the two scales are assumed to be uncorrelated. This assumption, however, is difficult to justify unless the background error is homogeneous and isotropic (Boer, 1983; Berre, 2000).

In the data assimilation experiments of which results will be presented in Section 4, Equation (8) is substituted into $p^f$ in Eqs. (12) and (13) for consistency. We also conducted data assimilation experiments in which $p^f = \langle \Delta x_f (\Delta x_f)^T \rangle$ was substituted for $p^f$ in these equations, and no significant differences were found in the accuracy of the analysis. It implies that essential to the PV-based EnKF is the neglect of the analysis increments proportional to $\langle \Delta x^f_f (\Delta x^f_f)^T \rangle$ or $\langle \Delta x^f_u (\Delta x^f_u)^T \rangle$ in the analysis equation of the Kalman filter. Equations (12) and (13) are suitable for the stochastic EnKF (Evensen, 1994; Burgers et al., 1998) and the serial ensemble square root filter (Serial EnSRF; Whitaker and Hamill, 2002). To apply Eq. (11) to the local ensemble transform Kalman filter (LETKF; Hunt et al., 2007), which uses the analysis covariance matrix $P^a$ in computing the Kalman gain, the inverse matrix $(R + HP^f H^T)^{-1}$ in Eqs. (12) and (13) is to be replaced by $R^{-1}(R - HP^a H^T)R^{-1}$. A problem with the LETKF is that it is difficult to compute $P^a$ under the approximation of Eq. (8). The above result, however, suggests that the $P^a$ computed in the conventional way of the LETKF can be used in calculating $K_b$ and $K_u$, but the LETKF is not perused in this paper.
The approximation to the Kalman gain given by Eqs. (11)-(13) has another benefit; different covariance localizations can be employed in computing $K_b$ and $K_u$. Since the unbalanced part is usually of smaller scales, a smaller localization radius may be appropriate for $K_u$. In addition, there are more degrees of freedom in the unbalanced part than in the balanced part. It implies more sampling errors and the need for stronger localization. Multi-scale localization methods like this have been proposed to cope with a wide range of scales in ensemble-based data assimilation (Zhang et al. 2009; Miyoshi and Kondo, 2013; Buehner and Shlyaeva, 2015). A crucial issue with covariance localization is its deleterious effect on dynamical balance (e.g., Lorenc, 2003; Houtekamer and Mitchell, 2005; Kepert, 2009; Bannister, 2015). In the PV-based EnKF, the balanced and unbalanced parts are handled separately in covariance localization and the balance issue of localization could be appropriately addressed. This is a distinct feature of the proposed method. Although further discussion will be given in Section 5, it is beyond the scope of the present paper to investigate the multi-scale covariance localization of the PV-based EnKF.

In the stochastic EnKF and the serial EnSRF, observational data are serially assimilated in batches or one by one. The analysis $x^a$ obtained by assimilating a batch of observations is used as the forecast $x^f$ to compute the Kalman gain for assimilating the next batch of observations. The application of PV inversion to obtain $K_b$ and $K_u$ at every step of serial assimilation is too costly. We can approximately decompose $x^a$ without applying PV inversion. To show this, a balance operator $B$ that extracts the balance part $x_b$ from the
dynamical state variables $\mathbf{x}$ is introduced: that is, $\mathbf{x}_{b} = \mathbf{B}(\mathbf{x})$. This operator consists of the spatial smoothing, computation of PV, and PV inversion. Assume that the norm of the analysis increment is much smaller than the norm of $\mathbf{x}^f$. Then, using the analysis equation of the Kalman filter and Eqs. (8)-(13), we obtain the following approximations:

$$
\mathbf{x}_{b}^a \approx \mathbf{B}\left[\mathbf{x}^f + \mathbf{K}\left(\mathbf{y}^o - \mathbf{H}(\mathbf{x}^f)\right)\right]
$$

$$
\approx \mathbf{B}(\mathbf{x}^f) + \left.\frac{\partial \mathbf{B}}{\partial \mathbf{x}}\right|_{\mathbf{x}^f} \mathbf{K}\left(\mathbf{y}^o - \mathbf{H}(\mathbf{x}^f)\right)
$$

$$
\approx \mathbf{x}_{b}^f + \mathbf{K}_{b}\left(\mathbf{y}^o - \mathbf{H}(\mathbf{x}^f)\right),
$$

(14)

$$
\mathbf{x}_{u}^a \approx \mathbf{x}_{u}^f + \mathbf{K}_{u}\left(\mathbf{y}^o - \mathbf{H}(\mathbf{x}^f)\right),
$$

(15)

where $\mathbf{x}_{b}^f$ and $\mathbf{x}_{u}^f$ are the balanced and unbalanced parts of $\mathbf{x}^f$, $\mathbf{y}^o$ is the observation vector, and $\mathbf{H}$ is the observation operator. The tangent linear approximation, $\Delta \mathbf{x}_{b}^a \approx \partial \mathbf{B}/\partial \mathbf{x}\bigg|_{\mathbf{x}^f} \Delta \mathbf{x}$, is used to derive the third line of Eq. (14), assuming the smoothness of the operator $\mathbf{B}$. Equation (15) follows from Eqs. (11) and (14). If the quasi-geostrophic inversion is adopted, then the balance operator is linear and Eqs. (14) and (15) are exact under the assumption of Eq. (8). However, if covariance localization is applied to $\mathbf{P}_{b}^f$ and $\mathbf{P}_{u}^f$, these equations become less accurate. The $\mathbf{x}_{b}^a$ and $\mathbf{x}_{u}^a$ obtained by Eqs. (14) and (15) can be used to compute $\mathbf{K}_{b}$ and $\mathbf{K}_{u}$ for assimilating the next batch of observations in serial assimilation, after applying the mass adjustment if necessary. This procedure is much superior to the application of PV inversion in saving the computational cost.
3. Design of experiments

We conduct data assimilation experiments with a shallow water model to examine the performance of the PV-based EnKF compared to a conventional EnKF by using twin experiments in a perfect model scenario. The only difference between the two EnKFs is that the approximation to the Kalman gain given by Eqs. (11)-(13) is not employed in the conventional EnKF.

3.1 Numerical model

The shallow water equations are the simplest equations of fluid dynamics that include both slow and fast variables. The governing equations of a shallow water model on an $f$-plane, which are nondimensionalized in terms of a characteristic length scale $L$ and a characteristic velocity scale $U$, are written as

\begin{align}
\frac{\partial \zeta}{\partial t} + R_o^{-1} D &= -\nabla \cdot \zeta v + \nu \nabla^2 \zeta, \\
\frac{\partial \theta}{\partial t} + R_o^{-1} \nabla^2 h - R_o^{-1} \zeta &= -\nabla \cdot (v \cdot \nabla) v + \nu \nabla^2 D, \\
\frac{\partial h}{\partial t} + R_o^{-1} \gamma^2 D &= -\nabla \cdot h v + \nu \nabla^2 h,
\end{align}

where $v = (u, v)$ is horizontal velocity, $\zeta$ is relative vorticity, $D$ is divergence, and $h$ is the departure of height from its reference value. The height departure $h$ is the mass variable referred to in Subsection 2.2. The Rossby number, $R_o = U/f L$, and the rotational Froude number, $\gamma = fL/\sqrt{gH}$, are introduced, where $f$ is the Coriolis parameter, $g$ is the
acceleration of gravity and $H$ is the reference value of height. The PV of the reference state is given by $1/R_o$, and the departure of PV from the reference value, which is hereafter referred to as the PV anomaly, is given by

$$\Delta P = \frac{\zeta - \gamma^2 h}{1 + R_o \gamma^2 h}$$

(19)

The Rossby deformation radius is adopted as the characteristic length $L$ and the Rossby number is set to 0.5. Then $\gamma = 1$ and the Froude number, $F_r = U/\sqrt{gH}$, is 0.5, and the nondimensional value of the reference height is 2.

The computational domain is doubly periodic, spanning the range $(-\pi, \pi)$ in the $x$ and $y$ directions. Inertial oscillations are possible under this periodic boundary condition, but they are removed for simplicity. A spectral method with the double Fourier expansion is adopted for spatial discretization, and a transform method is used for calculating the spectral coefficients of nonlinear terms. The number of grid points adopted for the transform method is $64 \times 64$. The highest resolved wavenumber is set to 21 in each direction to avoid aliasing, and it is denoted by $k_{\text{max}}$. Thus, the degrees of freedom of the model is $3[4 k_{\text{max}}(k_{\text{max}} + 1) + 1] - 2 = 5545$, since the area-averaged values of $\zeta$ and $D$ vanish. The subtracted two degrees of freedom belong to inertial oscillations removed from the model.

Time integration is conducted by the leapfrog scheme in combination with the Euler backward scheme, which is inserted at every ten time-steps to suppress computational modes. The size of time-step is set to 0.01. The value of numerical dissipation $\nu$ is
determined by \( v = |\Delta P|_{\text{max}} / k_{\text{max}}^2 \) according to Dritschel et al. (1999), where \( |\Delta P|_{\text{max}} \) is the maximum value of \( |\Delta P| \) in the domain. In the present study, \( |\Delta P|_{\text{max}} \) is taken to be a constant, which is estimated from the true state at the initial time used in the data assimilation experiments. This shallow water model is used for PV inversion, true state generation, and data assimilation with the same resolution.

3.2 PV inversion

The state variables of the shallow water model, \( \nu \) and \( h \), are partitioned into the balanced and unbalanced parts by applying PV inversion using the first-order direct inversion of McIntyre and Norton (2000). The quasi-geostrophic inversion is also employed in some cases for comparison. The quasi-geostrophic PV anomaly for the latter inversion is given by Eq. (19) with \( R \) set to zero. We conducted a preliminary data assimilation experiment using the second-order direct inversion of McIntyre and Norton (2000), which derived the balanced part of divergence from the PV anomaly, but we obtained no improvement in the accuracy of the analysis despite more computational cost and more failures in convergence. Note that there is no arbitrariness of boundary conditions in PV inversion in cases with the periodic boundary condition.

The state variables of a forecast ensemble are spatially smoothed for computing the large-scale PV anomaly. The smoother used in the present study is similar to the one proposed by Hoskins (1980). The state variables in spectral space with two-dimensional
wavenumber \((k, l)\) are multiplied by \(\exp[-K(k^2 + l^2)^2]\), where the coefficient \(K\) is set to 
\[2 \log 10/(2k_{max}^2)^2\] such that the amplitude of state variables with the wavenumber 
\((k_{max}, l_{max})\) becomes two orders of magnitude smaller than the original size.

For the first-order direct inversion, the balanced velocity \(v_b\) and balanced height 
depture \(h_b\) are obtained by solving the following equations:

\[\zeta_b - \gamma^2 h_b = \Delta P(1 + R \gamma^2 h_b), \quad (20)\]

\[\nabla^2 h_b = \zeta_b + 2R \left[ \frac{\partial^2 \psi_b}{\partial x^2} \frac{\partial^2 \psi_b}{\partial y^2} - \left( \frac{\partial^2 \psi_b}{\partial x \partial y} \right)^2 \right], \quad (21)\]

\[\zeta_b = \nabla^2 \psi_b, \quad (22)\]

where \(\zeta_b\) and \(\psi_b\) are the balanced parts of relative vorticity and a stream function, respectively. Equation (20) is derived from Eqs. (1) and (19), and Eq. (21) is the Bolin-Charney balance condition corresponding to Eq. (3). The balanced velocity \(v_b\) is computed from \(\psi_b\) by using Eq. (4). Since Eq. (21) is nonlinear, an iterative algorithm is taken. See Appendix for details of the iterative algorithm. If the terms multiplied by \(R\) are removed in Eqs. (20)-(22), they are reduced to the equations for the quasi-geostrophic inversion. After solving Eqs. (20)-(22), the mass adjustment is applied to the balanced and unbalanced height departure, \(h_b\) and \(h_u\), by using the equations corresponding to Eqs. (6) and (7).

There are a couple of issues on PV inversion. The first issue is the uniqueness of solutions. For the first-order direct inversion, the following elliptic equation for \(h_b\) is
obtained from Eqs. (20)-(22):

\[
\left[ \varphi^2 - R_o y^2 \left( \frac{1}{R_o} + \Delta P \right) \right] h = \Delta P + 2R \left[ \frac{\partial^2 \psi_b}{\partial x^2} \frac{\partial^2 \psi_b}{\partial y^2} - \left( \frac{\partial^2 \psi_b}{\partial x \partial y} \right)^2 \right].
\]  

(23)

We can easily prove that the boundary value problem, Eq. (23) under the periodic boundary condition, has a unique solution if PV, which is given by \(1/R_o + \Delta P\), is not negative and not identically zero in the domain. In the present study, if PV is negative somewhere in the domain, the negative values are replaced with zeros before applying PV inversion. An additional data assimilation experiment, however, showed that even without this correction of PV, convergence failure did not occur and the accuracy of the analysis was almost the same as that of the analysis with the correction.

The other issue is the mass conservation. For the first-order direct inversion, the following equation can be derived from Eqs. (19) and (20) under the periodic boundary condition (noting that \(\bar{\zeta}\) and \(\bar{\zeta}_b\) are zero):

\[
\bar{h} - \bar{h}_b = -R_o \Delta P (\bar{h} - \bar{h}_b).
\]

(24)

There is no guarantee that \(\bar{h}_b\) is equal to \(\bar{h}\) unless \(\Delta P\) is a constant in the domain. If \(\Delta P\) is a constant and the sufficient condition for the uniqueness of solutions mentioned above is satisfied, the equation \(\bar{h}_b = \bar{h}\) follows from Eq. (24). Note that if the boundary condition is of a Dirichlet type, we can adjust the boundary values of \(h_b\) iteratively such that the mass conservation holds. For the quasi-geostrophic inversion, the right-hand side of Eq. (24) vanishes, so that mass is unconditionally conserved.
3.3 True state

The initial condition of the true state adopted in the experiments is a superposition of a balanced jet flow and an unbalanced height anomaly. According to Dritschel et al. (1999), the balanced jet flow in the initial condition is specified by prescribing the PV anomaly as follows:

\[
\Delta P = \begin{cases} 
\frac{\Delta P}{\Delta P_0} + \Delta P_0 \text{sign}(y)(\alpha - |y| - \alpha) & (|y| < 2\alpha) \\
\frac{\Delta P}{\Delta P_0} & (2\alpha \leq |y| \leq \pi)
\end{cases}
\]

where \(2\alpha\) is the distance between the maximum and minimum of PV anomaly, \(2\alpha \Delta P_0\) is the difference between the maximum and minimum values of PV anomaly, and \(\Delta P\) is the area-averaged PV anomaly which is determined by the requirement that the area averaged height departure vanishes. The parameter values taken in the experiments are \(\Delta P_0 = 4\), \(\alpha = 0.5\). Since \(R_o\) is set to 0.5, the value of reference potential vorticity is 2 and \(\Delta P = 0.336\), so that the PV of the balanced jet flow is positive everywhere in the domain. As the gradient of PV anomaly changes sign in the domain, the jet flow satisfies a necessary condition for barotropic instability to quasi-geostrophic perturbations. The stream function \(\psi\) and height departure \(h\) are computed from \(\Delta P\) by applying the first-order direct inversion.

The unbalanced height anomaly in the initial condition is given by

\[
h(x, y) = \tilde{h} + h_0 \exp \left[ -\frac{(x-x_0)^2+(y-y_0)^2}{\alpha^2} \right],
\]

(25)
where $h_0$ is the magnitude of height anomaly, $(x_0, y_0)$ is the location of the maximum anomaly, $d$ is the horizontal size of the anomaly, and $\tilde{H}$ is determined by the requirement that the area-averaged height anomaly vanishes. The parameter values taken in the experiments are $h_0 = 1$, $d = 0.5$, and $(x_0, y_0) = (0, 1.5)$. The velocity associated with the height anomaly is set to zero, so that the height anomaly is unbalanced.

The initial condition of the true state is constructed by superposing the velocity and height departure fields of the balanced jet flow and the unbalanced height anomaly mentioned above. The PV of the initial condition is positive everywhere in the domain.

Figure 1 displays the time evolution of height departure and velocity for a period of $0 \leq t \leq 200$. The balanced jet flow and the unbalanced height anomaly at the initial time ($t = 0$) almost do not overlap. After $t = 0$ inertial gravity waves start to propagate from the unbalanced height anomaly to the balanced jet flow, and the latter is broken into several vortices by $t = 25$. A pair of cyclonic and anticyclonic vortices prevails in the last quarter period. Figure 2 displays the evolution of divergence for the same period, which primarily represents the unbalanced part. The white dots in Fig. 2a indicate the positions of observations that will be mentioned in the next subsection. Though there is no divergence at $t = 0$, small-scale divergences are immediately generated by inertial gravity waves.

Since the Euler backward scheme is used at every ten time-steps in the time integration, inertial gravity waves with smaller scales gradually decay with time (e.g., Kalnay, 2003). Note that the spatial scale of divergence around after $t = 50$ is not very different from that
of height departure. Thus, the spatial scales of the balanced and unbalanced part are not very different except for the first quarter period in the experiments.

3.4 Data assimilation

The stochastic EnKF is chosen as the method of data assimilation for an illustration, even if it may not be the first choice in practice because of sampling errors caused by perturbed observations (e.g., Bowler et al., 2013). Sumitomo (2018) examines the performance of the PV-based EnKF proposed in the present paper for the serial EnSRF using a shallow water model and demonstrates its benefits similar to those presented in Subsection 4.3. Data assimilation is conducted for a period of $0 < t \leq 200$ with ensemble sizes of 25, 50, 100, and 1 000. To reduce sampling errors in the results, data assimilation experiments are repeated ten times using different random number sequences for perturbations to the observations and to the initial conditions of a forecast ensemble.

Since velocity data are usually more useful than height data for the analysis of small scales in the atmosphere, velocity observations are assimilated in the present study for an illustration. Observations are the $x$ and $y$ components of velocity for a period of $0 < t \leq 200$. They are generated by adding random observation errors to the true state. The observation errors are independent random draws from the Gaussian distribution with the mean 0. The observations are available at a space interval of $\pi/4$ in each direction at the positions indicated in Fig. 2a, so that the number of observations in the domain is 128.
To reduce the computational cost, observational data are serially assimilated in batches with 16 observations at a time in the stochastic EnKF. The time interval of observations is set to 1. As the characteristic time scale of the unbalanced part is about 1 (see Fig. 3 below), the observations are not sufficient to accurately analyze the unbalanced part. The observation errors of the velocity components are uncorrelated, and their standard deviation is set to 0.05, which is one order of magnitude smaller than the characteristic size of the velocity of the true state (see Fig. 1). We also conducted data assimilation experiments with height observations and obtained qualitatively similar results to those given in Section 4 except for the benefits of the mass adjustment, which was small as may be expected. The results of these experiments are not shown in this paper to keep it concise.

The analysis variables are velocity $\mathbf{v}$ and height departure $h$. Multiplicative covariance inflation and covariance localization are applied to suppress the adverse effects of sampling errors. The forecast error covariance matrix is multiplied by a factor $\rho$, which is referred to as the inflation factor. Covariance localization is implemented as a Schur product of the ensemble-sampled forecast error covariance matrix for $\mathbf{v}$ and $h$ and a correlation matrix. The correlation function defined by Eq. (4.10) of Gaspari and Cohn (1999) is adopted for the correlation matrix. Their parameter $c$ is referred to as the localization radius in the present paper, at which radius the correlation coefficient decreases to 5/24. The covariance inflation factor and localization radius do not depend on
space and time. In the present study, the same localization radius is used for the Kalman
gains of the balanced and unbalanced parts, as we did not find the benefits of using
different localization radii in preliminary data assimilation experiments. This is probably
because the spatial scale of the unbalanced part is not very different from that of the
balanced part except for the first quarter period, as mentioned in the previous subsection.

The initial condition of the forecast state is prepared by shifting the position of the
balanced jet flow of the true state by -1 and that of the local height anomaly of the true state
by 1 in the $y$ direction. The initial conditions of the forecast ensemble are prepared by
independently shifting the balanced jet flow and the local height anomaly of the forecast
state in the $x$ and $y$ directions by random numbers according to the standard normal
distribution. As a result, there are no perturbations to divergence in the initial conditions of
the forecast ensemble. Those initial perturbations help to shorten the spin-up stage in data
assimilation. Note that although the area-averaged height departure $\bar{h}$ vanishes in all
forecast members at the first assimilation time ($t = 1$), the analyzed value of $\bar{h}$ at that time
deviates from zero due to the use of covariance localization.

For the PV-based EnKF, we apply the following procedures in this order to the state
variables of each member of a forecast ensemble: spatial smoothing, computation of PV,
PV inversion, and mass adjustment. Then we obtain the ensembles of $\Delta x^f_b$ and $\Delta x^f_{u}$, and
compute $p^f_b$ and $p^f_{u}$ according to Eqs. (9)-(10). The approximate Kalman gain given by
Eqs. (11)-(13) is used to assimilate the first batch of observations in the stochastic EnKF.
When assimilating the remaining batches of observations, Equations (14) and (15) are employed for the decomposition of $x^a$ instead of applying PV inversion. We found from a preliminary data assimilation experiment of the PV-based EnKF that despite the use of covariance localization the balanced and unbalanced parts of $x^a$ computed by Eqs. (14) and (15) were almost the same as those obtained by applying PV inversion to $x^e$. We found from another preliminary experiment that the analyzed value of $h^a$ remained nearly zero during serial assimilation, so that the mass adjustment is not applied to $x^c$ and $x^c_0$ when Eqs. (14) and (15) are used. In addition, we experienced occasional failures in convergence of the first-order direct inversion at the first assimilation time ($t = 1$) in preliminary data assimilation experiments. The conventional EnKF is therefore employed to compute the analysis at $t = 1$ in data assimilation experiments of the PV-based EnKF.

4. Results of experiments

In this section the root mean square error (RMSE) of analysis and the spread of analysis are defined as

$$\text{RMSE} = \sqrt{(u-\bar{u})^2 + (v-\bar{v})^2 + \gamma^2(h-\bar{h})^2},$$

$$\text{Spread} = \sqrt{\langle (u-\langle u \rangle)^2 \rangle + \langle (v-\langle v \rangle)^2 \rangle + \gamma^2 \langle (h-\langle h \rangle)^2 \rangle},$$

respectively, where a superscript $t$ denotes the true state. Those definitions are based on the perturbation energy of the shallow water equations. Results presented below for the
PV-based are ones obtained by using the first-order direct inversion unless otherwise stated.

4.1 PV inversion of true state

At first, the performance of the first-order direct version and quasi-geostrophic inversion in the decomposition of state variables are examined by applying them to the true state used in the experiments. Figure 3 shows the time series of the balanced and unbalanced parts of height departure of the true state at the center of the domain. The top panel is for the quasi-geostrophic inversion, and the bottom one is for the first-order direct inversion. They are computed at a time interval of 0.1 with the spatial smoothing. The latter inversion converges within ten iterations. Roughly speaking, both PV inversions successfully decompose state variables. The balanced part evolves very slowly compared to the unbalanced part. Oscillations of the balanced height with a period of about 10 after $t = 160$ are caused by a clockwise rotation of the distorted anticyclonic vortex seen in Figs. 1e and 1f. It is also found from Fig. 3 that the time scale of the unbalanced part is about 1, so that the observation interval of 1 used in the experiments is insufficient for accurate analysis of the unbalanced part.

Further inspection of Fig. 3, however, indicates that the unbalanced part derived by the quasi-geostrophic inversion contains low-frequency variations, which are negatively correlated with the balanced part. The correlation coefficient between the balanced and
unbalanced parts of height departure at the center of the domain is -0.499 for the 
quasi-geostrophic inversion, while the corresponding correlation coefficient for the 
first-order direct inversion is -0.010. Therefore, the latter inversion is superior to the former 
in the decomposition of state variables. A problem with the first-order direct inversion is that 
the balanced part contains small-amplitude fluctuations with the same time scale as the 
unbalanced part and that they are more prominent than those of the quasi-geostrophic 
inversion. Their amplitude, however, is one order of magnitude smaller than that of the 
unbalanced part, so that the influence of these high-frequency fluctuations may be 
negligible.

When the shallow water model is integrated from a balanced flow of the true state 
derived by the first-order direct inversion, state variables remain close to the balanced part 
obtained by this inversion up to several tens of time, but without high-frequency fluctuations 
mentioned above. On the other hand, when a balanced flow derived by the 
quasi-geostrophic inversion is used as an initial condition, state variables become closer to 
the balanced part by the first-order direct inversion rather than that by the quasi-geostrophic 
inversion, and contain high-frequency fluctuations with a larger amplitude than those seen 
in the balanced part in Fig. 3b (not shown). This difference in the results of time integration 
between the two inversions is similar to the difference between the linear and nonlinear 
normal mode initializations in NWP (e.g., Kalnay, 2003). This result also indicates that the 
balanced part derived by the first-order direct inversion is more balanced than that by the
quasi-geostrophic inversion.

As mentioned in Subsection 3.2, the first-order direct inversion does not conserve mass for the shallow water equations with the periodic boundary condition. Figure 4 shows the time sequence of the balanced height departure of the true state averaged over the domain. It is computed at a time interval of 0.1 with the spatial smoothing as in Fig. 3. Although $\bar{h_b}$ fluctuates around zero with the same time scale as that of the unbalanced part, its amplitude is two orders of magnitude smaller than the characteristic amplitude of $h_b$ (see Fig. 3) and its time-averaged value over the whole period is only $-2 \times 10^{-5}$. Therefore, the mass conservation is not a serious issue.

4.2. Effects of mass adjustment and spatial smoothing

The necessity of the mass adjustment is demonstrated first. Figure 5 compares the time series of the RMSEs of analysis and the biases of the area-averaged height analysis with and without the mass adjustment. The corresponding time series for the conventional EnKF is also shown. The ensemble size is 50, and the covariance inflation factor and localization radius are optimized (see Fig. 8b below); they are set to 1.05 and 15 grid points (1 grid point is $\pi/32$ in distance), respectively. The corresponding result for the quasi-geostrophic inversion is not very different from Figs. 5a and 5b (not shown). As mentioned in Subsection 3.4, the data assimilation experiments are repeated ten times using different random number sequences. All of these results are presented in the figure in different
colors, in which the results obtained by using the same random number sequence are plotted in the same color. It is found from Fig. 5a that when the mass adjustment is not applied the bias of the area-averaged height analysis strongly depends on random number sequences, and that in some cases the magnitude of the bias becomes over half of the RMSE of analysis in the latter half period. Figure 5b shows that the mass adjustment effectively eliminates those problems, and the accuracy of the analysis is improved. This result indirectly justifies the assumption of Eq. (5); the forecast error covariances between the balanced horizontal velocity and the area-averaged value of the balanced mass variable can be ignored. The corresponding result of the conventional EnKF is displayed in Fig. 5c, and a comparison with Fig. 5b indicates that the PV-based EnKF with the mass adjustment is better than the conventional EnKF in terms of the RMSE of analysis and the bias of the area-averaged height analysis. The RMSEs averaged over the random number sequences for this case are compared in Fig. 9b below.

Figure 6 compares the RMSEs with and without the mass adjustment for the ensemble sizes of 25, 50, 100, and 1 000. They are the averages over the ten random number sequences. The covariance inflation factors and localization radii are optimized for each ensemble size (see Fig. 8 below). It is found from Fig. 6 that the smaller the ensemble size is, the larger the benefits of the mass adjustment are. Even when the ensemble size is 1 000, in which the covariance localization is not applied, the mass adjustment still has a marginally positive impact. Figure 6 also shows a clear tendency that the spin-up period of
data assimilation becomes shorter with an increase of the ensemble size.

Figure 7 compares the RMSEs with and without the spatial smoothing in the same format as that of Fig. 6. If the coefficient $K$ of the spatial smoothing defined in Subsection 3.2 is doubled from the value used in Fig. 7, the accuracy of the analysis is slightly degraded (not shown). It suggests that there is an optimal degree of the spatial smoothing. The benefits of the spatial smoothing are seen in most cases, although it is not large. An exception is for the ensemble size of 1000 for a period from $t = 20$ to $t = 70$. This adverse effect of the spatial smoothing is an unexpected result, and a sure explanation has not been found. The following is a speculation. If PV anomaly for PV inversion is computed from spatially smoothed state variables, the unbalanced part contains a part of the state variables of which dynamics is governed by the conservation of PV. If the ensemble size is sufficiently large, weak error correlations between the balanced and unbalanced parts are correctly computed, and it may become necessary to appropriately handle the above mixed nature of the unbalanced part. The neglect of error correlations between the balanced and unbalanced parts would make this process difficult when the unbalanced part is very active.

4.3 Accuracy of analysis

In this subsection, the accuracy of the analysis is compared between the PV-based and conventional EnKFs. Figure 8 displays the RMSEs averaged over the period from $t = 25$ to $t = 125$ and over the ten random number sequences as a function of the covariance
inflation factor and localization radius for the ensemble size of 25, 50, 100 and 1 000. The spatial smoothing for PV inversion is applied to all the ensemble sizes, although it has a detrimental effect for the ensemble size of 1 000 as mentioned in the previous subsection. The above time-mean period is chosen to avoid strong spin-up in an early period and to see the benefits of the PV-based EnKF when the unbalanced part is relatively active. Solid lines are for the PV-based EnKF and broken lines are for the conventional EnKF. They are plotted against the inflation factor at an interval of 0.025 and the localization radius at an interval of 5 grid points except for the ensemble size of 1 000, in which the RMSEs are plotted at an interval of 0.01 for the inflation factor and covariance localization is not applied. If covariance localization with the localization radius of 100 grid points is used in this case, the RMSEs of both EnKFs increase slightly and almost equally. For the ensemble size of 25, when the localization radius is set to 5 grid points and the inflation factor is larger than 1.075, the PV-based EnKF fails due to the convergence failure of PV inversion for most or all of the random number sequences, whereas with the conventional EnKF no difficulties arise. This is probably because the PV-based EnKF applies covariance localization to the balanced and unbalanced parts of state variables separately and, therefore, its deleterious effect on balance may be more serious than the conventional EnKF. For the ensemble size of 100, when the localization radius is set to 30 grid points, both EnKFs fail at the first assimilation time ($t = 1$) for a couple of random number sequences. Hence, the results for this localization radius are not presented in Fig. 8c.
The minimum of RMSE for each EnKF with each ensemble size is indicated by a light blue marker in Fig. 8. The values of the inflation factor and localization radius that yield the minimum are referred to as the optimum values in the present paper, although the minimum shown in the figure may not be the exact minimum. It is found from this figure that the smaller the ensemble size is, the greater the degree of improvement over the conventional EnKF is. The computational time of the PV-based EnKF is about 10 % more than that of the conventional EnKF in the experiments. It is therefore concluded from Figs. 8a-8c that the PV-based EnKF improves the accuracy of the analysis much more efficiently than increasing the ensemble size. If a low-resolution model is used for PV inversion, further reduction of computational cost would be obtained. When the ensemble size is increased to 1,000, however, the conventional EnKF outperforms the PV-based EnKF. This result does not change even if the spatial smoothing is not applied, as will be mentioned below in this subsection. If the ensemble size is sufficiently large, weak error correlations between the balanced and unbalanced parts can be correctly computed, and these weak correlations are effectively used in the conventional EnKF. Hence, the above result is an expected one.

Another result found from Fig. 8 is that the optimal inflation factor of the PV-based EnKF tends to be smaller than that of the conventional EnKF. This result suggests that sampling errors in the former EnKF are smaller than those in the latter EnKF.

The time series of RMSEs and spreads of the PV-based and conventional EnKFs using the optimal inflation factors and localization radii are compared in Fig. 9 for ensemble sizes...
of 25, 50, 100 and 1 000. Dark blue lines are for the PV-based EnKF using the first-order direct inversion, light blue lines are for the PV-based EnKF using the quasi-geostrophic inversion, and light green lines are for the conventional EnKF. The plotted RMSEs and spreads are averaged values over the ten random number sequences. Note that the optimum inflation factor is different between the two EnKFs for ensemble sizes of 25 and 100, while the optimum value of localization radius is the same for all ensemble sizes.

It is found from Fig. 9 that for the ensemble sizes of 25, 50 and 100 the PV-based EnKF is superior to the conventional EnKF during most of the period in terms of the accuracy of the analysis. A comparison with Fig. 6 indicates that the mass adjustment is necessary for the PV-based EnKF to outperform the conventional EnKF. For the ensemble size of 1 000, it is up to about \( t = 70 \) that the conventional EnKF outperforms the PV-based EnKF. A comparison between Fig. 7 and Fig. 9d shows that the RMSE without the spatial smoothing is also larger than that of the conventional EnKF. The unbalanced flow is very active in this early period (see Fig. 2), so that the error covariances between the balanced and unbalanced parts may be too large to be ignored. The RMSEs of the PV-based EnKF for the two inversion methods are almost the same except for the ensemble size of 50, in which the accuracy of the analysis for the first-order direct inversion is slightly better than the quasi-geostrophic inversion. Therefore, although the first-order direct inversion is superior to the quasi-geostrophic inversion in the decomposition of state variables, the benefits of the former inversion over the latter in data assimilation are marginal.
There is a tendency that the spread of the PV-based EnKF is larger than that of the conventional EnKF when the same inflation factor and localization radius are used (Figs. 9b and 9d). As already mentioned in this subsection, the optimal inflation factor of the PV-based EnKF tends to be smaller than that of the conventional EnKF. Since the spread tends to increase with an increase of the inflation factor, the larger spreads of the PV-based EnKF for the same inflation factor is an expected result.

5. Summary and discussion

A multi-scale data assimilation method for the EnKF is proposed for atmospheric models in cases with insufficient observations of fast variables. The method is based on the conservation and invertibility of PV. It consists of three components: (i) decomposition of the dynamical state variables in the free atmosphere of forecast ensemble members into the balanced and unbalanced parts by applying PV inversion to the PV anomalies computed from spatially smoothed state variables, (ii) adjustment of mass variables of the two parts to remove additional sampling errors introduced by this decomposition, and (iii) approximation to the Kalman gain by ignoring forecast error covariances between those parts. This approximation makes it possible to apply different covariance localizations to each part, even though the same localization is applied to each in this study.

The performance of the proposed data assimilation method is demonstrated with a shallow water model through twin experiments in a perfect model scenario, in which
observations of velocity components are assimilated. Results using the same localization radius for the two parts show that the PV-based EnKF is superior in the accuracy of the analysis to the conventional EnKF unless the ensemble size is sufficiently large. It is demonstrated that the mass adjustment is necessary for the PV-based EnKF to outperform the conventional EnKF and that the spatial smoothing for PV inversion has a mostly positive impact. The benefits of using the first-order direct inversion over the quasi-geostrophic inversion are marginal in the experiments.

In the present study, the PV-based EnKF is applied to the shallow water model. Although the demonstration of the proposed method is based on a perfect model scenario, a similar performance is expected in realistic applications with imperfect models, since current atmospheric models can simulate a balanced flow very well. In implementing this method to a three-dimensional high-resolution atmospheric model, the decomposition of state variables should be conducted to the dynamical state variables in the free atmosphere only.

In other words, the ignored part of a forecast error covariance matrix is the covariances between balanced and unbalanced flows in the free atmosphere. The state variables in the planetary boundary layer and those of water substance including water vapor are handled in the same way as in the conventional EnKF. Then, for instance, error correlations associated with interactions between cumulus convection and the planetary boundary layer are intact. Error correlations between the small-scale stationary flow in the planetary boundary layer induced by the inhomogeneity of land surfaces and the large-scale flow are
retained. Another point to be noted is that we can conduct PV inversion by using a hydrostatic model with a lower resolution than the model used for data assimilation. Thereby the computational cost of PV inversion can be reduced, and the difficulty in solving the ill-conditioned differential equations is alleviated. For the operational implementation, PV inversion methods based on initialization such as that proposed by Arbogast et al. (2008) may be recommended, since those methods do not need to solve the ill-conditioned differential equations for state variables.

An important issue not addressed in this paper is the performance of multi-scale covariance localization of the PV-based EnKF. As mentioned in Subsection 3.4, the experimental setup used in the present study may not be appropriate to investigate the benefits of applying different localization radii to the balanced and unbalanced parts of state variables. Since the PV-based EnKF explicitly handles the balanced part of state variables, it is possible to implement different localization methods to the two parts. According to Kepert (2009), the localization in stream function-velocity potential space better preserves the balances contained within the unlocalized covariances than the localization in velocity component space used in the present study. However, the former localization method is not appropriate for the localization of the unbalanced part, since it is far from in geostrophic balance. A problem with the PV-based EnKF in implementing the localization in stream function-velocity potential space is that a difficulty would arise from the nonlinearity of the Bolin-Charney balance of Eq. (3). We can avoid this problem by conducting covariance
localization for the balanced part in PV space, but more computational cost is needed for additional PV inversion. To investigate the benefits of multi-scale covariance localization of the PV-based EnKF, data assimilation experiments using a high-resolution three-dimensional atmospheric model may be appropriate.

Although we do not suppose that the state variable of water vapor is decomposed into the balanced and unbalanced parts, such decomposition, if possible, would contribute to further reduction of spurious error correlations in the PV-based EnKF. A problem is that a plausible balance condition does not exist for water vapor. Generalizations of PV for a moist atmosphere have been proposed by several authors (Schubert et al., 2001; Gao et al., 2004; Peng et al., 2013; Marquet, 2014). According to them, a moist PV defined by using virtual potential temperature has an invertibility principle, but the partition of the total density into the densities of dry air and water substance is not possible from knowledge of the moist PV only. An acceptable “balanced” part of water vapor would be the part advected by the balanced flow. To explore the possibility of estimating this part of water vapor may be a topic of future research.

As mentioned in the introduction, the application of PV to variational data assimilation has been investigated by several authors. In their methods, PV inversion is applied at the beginning of assimilation window of 4D-Var. The influence of approximations introduced at this point of time is gradually dominated over by dynamics in the course of time integration in the assimilation window. This is one of the benefits of 4D-Var over EnKF. For instance,
Tsuyuki (2014) shows that a non-Gaussian prior probability density function (PDF) that evolves according to the Liouville equation (e.g., Ehrendorfer, 1994) is implicitly used in 4D-Var under certain conditions even if a prior PDF at the beginning of the assimilation window is Gaussian. On the other hand, the approximations introduced by using PV inversion in EnKF are directly reflected in the analysis through the Kalman gain. It implies that more care may be needed for the application of PV to EnKF than that to 4D-Var. The necessity of the mass adjustment may be an example.

**Acknowledgments**

The author is grateful to Hiromu Seko, Takuya Kawabata, Sho Yokota and Daisuke Hotta of the Meteorological Research Institute for helpful discussion. The author would like to thank two anonymous reviewers for their careful reading and a lot of comments, which helped much to improve the manuscript. This study was partly supported by the Japan Society for the Promotion of Science (KAKENHI "Study on uncertainty of cumulonimbus initiation and development using particle filter"; JP17H02962).

**Appendix: Iterative algorithm of PV inversion**

For the first-order direct inversion, Equations (20)-(22) are solved by the following iterative algorithm, where $k$ is the count of iterations.

- For $k = 0$:
\[ \psi_k = h_k = 0. \]  
(A1)

- For \( k = 1, 2, 3, \ldots \):

\[ (\nabla^2 - \gamma^2)h_k = \Delta P (1 + R_o \gamma^2 h_{k-1}) + 2R_o \left[ \frac{\partial^2 \psi_{k-1}}{\partial x^2} \frac{\partial^2 \psi_{k-1}}{\partial y^2} - \left( \frac{\partial^2 \psi_{k-1}}{\partial x \partial y} \right)^2 \right], \]  
(A2)

\[ \nabla^2 \psi_k = \gamma^2 h_k + \Delta P (1 + R_o \gamma^2 h_k). \]  
(A3)

The solution to Eq. (A2) for \( h_k \) is substituted into the right-hand side of Eq. (A3). A spectral method is adopted to solve Eqs. (A2) and (A3) with the same spatial resolution as the shallow water model used in the data assimilation experiments.

- If \( \sqrt{(\nabla^2 \psi_k - \nabla^2 \psi_{k-1})^2} < \varepsilon \) and \( \sqrt{(h_k - h_{k-1})^2} < \varepsilon \), where an overbar denotes the area average in the domain and \( \varepsilon \) is a small positive number, the iteration is terminated. The criterion of convergence \( \varepsilon \) is set to \( 1 \times 10^{-5} \). Note that \( \nabla^2 \psi_k \) and \( h_k \) are of the order of unity since they are nondimensionalized by using characteristic scales.

The equations of the quasi-geostrophic inversion are obtained by setting \( R_o \) to zero in Eq. (A2) and by replacing Eq. (A3) with the equation \( \psi_k = h_k \), and the iteration is not necessary.

Note that the quasi-geostrophic PV is used for computing \( \Delta P \) in the quasi-geostrophic inversion. Equation (A1) guarantees that \( \psi_1 \) and \( h_1 \) of the first-order direct inversion are close to the solution of the quasi-geostrophic inversion if \( R_o \ll 1 \).

References

Arbogast, P., K. Maynard, and F. Crepin, 2008: Ertel potential vorticity inversion using a


Buehner, M., and A. Shlyaeva, 2015: Scale-dependent background-error covariance


List of Figures

Fig. 1. Height departure and velocity of the true state at (a) $t = 0$, (b) $t = 25$, (c) $t = 50$, (d) $t = 100$, (e) $t = 150$ and (f) $t = 200$. The scale of velocity vectors is the same as that of the axes.

Fig. 2. Divergence of the true state at (a) $t = 0$, (b) $t = 25$, (c) $t = 50$, (d) $t = 100$, (e) $t = 150$ and (f) $t = 200$. White dots in (a) indicate the positions of observations.

Fig. 3. Time sequences of the balanced (blue lines) and unbalanced (green lines) parts of the height departure at the center of the domain of the true state for (a) the quasi-geostrophic inversion and (b) the first-order direct inversion.

Fig. 4. Time sequence of the area average of balanced height departure of the true state for the first-order direct inversion.

Fig. 5. RMSEs of analysis (solid lines) and biases of area-averaged height analysis (broken lines) for (a) the PV-based EnKF without mass adjustment, (b) the PV-based EnKF with mass adjustment, and (c) the conventional EnKF. They are plotted for ten random number sequences. The lines are plotted in the same color for the same random number sequence. The ensemble size is 50, and the covariance inflation factor and localization radius are optimized.

Fig. 6. Comparison of RMSEs of analysis of the PV-based EnKF with (solid lines) and
without (broken lines) mass adjustment for ensemble sizes of 25, 50, 100, and 1 000. They are averaged values over ten random number sequences. The covariance inflation factor and localization length are optimized for each ensemble size. The RMSEs of the same ensemble size are plotted in the same color.

Fig. 7. Same as Fig. 6, except for the PV-based EnKF with (solid lines) and without (broken lines) spatial smoothing for PV inversion.

Fig. 8. Comparison of RMSEs of analysis of the PV-based EnKF (solid lines) and conventional EnKF (broken lines) for ensemble sizes of (a) 25, (b) 50, (c) 100, and (d) 1 000. The RMSEs are averaged values over a period from t=25 to t=125 and over ten random number sequences. They are plotted against the covariance inflation factor for three covariance localization radii, which are indicated by two-digit numbers in the legends (unit: grid points). The minimum values of RMSE for each EnKF are shown by light blue markers.

Fig. 9. Comparison of RMSEs (solid lines) and spreads (dashed lines) of analysis of the PV-based EnKF using the first-order direct inversion (dark blue), the PV-based EnKF using the quasi-geostrophic inversion (light blue) and the conventional EnKF (light green) for ensemble sizes of (a) 25, (b) 50, (c) 100, and (d) 1 000. They are averaged values over ten random number sequences. The covariance inflation factor and localization radius are optimized for each ensemble size.
Fig. 1. Height departure and velocity of the true state at (a) $t = 0$, (b) $t = 25$, (c) $t = 50$, (d) $t = 100$, (e) $t = 150$ and (f) $t = 200$. The scale of velocity vectors is the same as that of the axes.
Fig. 2. Divergence of the true state at (a) \( t = 0 \), (b) \( t = 25 \), (c) \( t = 50 \), (d) \( t = 100 \), (e) \( t = 150 \) and (f) \( t = 200 \). White dots in (a) indicate the positions of observations.
Fig. 3. Time sequences of the balanced (blue lines) and unbalanced (green lines) parts of the height departure at the center of the domain of the true state for (a) the quasi-geostrophic inversion and (b) the first-order direct inversion.
Fig. 4. Time sequence of the area average of balanced height departure of the true state for the first-order direct inversion.
Fig. 5. RMSEs of analysis (solid lines) and biases of area-averaged height analysis (broken lines) for (a) the PV-based EnKF without mass adjustment, (b) the PV-based EnKF with mass adjustment, and (c) the conventional EnKF. They are plotted for ten random number sequences. The lines are plotted in the same color for the same random number sequence. The ensemble size is 50, and the covariance inflation factor and localization radius are optimized.
Fig. 5. (Continued)
Fig. 6. Comparison of RMSEs of analysis of the PV-based EnKF with (solid lines) and without (broken lines) mass adjustment for ensemble sizes of 25, 50, 100, and 1000. They are averaged values over ten random number sequences. The covariance inflation factor and localization length are optimized for each ensemble size. The RMSEs of the same ensemble size are plotted in the same color.
Fig. 7. Same as Fig. 6, except for the PV-based EnKF with (solid lines) and without (broken lines) spatial smoothing for PV inversion.
Fig. 8. Comparison of RMSEs of analysis of the PV-based EnKF (solid lines) and conventional EnKF (broken lines) for ensemble sizes of (a) 25, (b) 50, (c) 100, and (d) 1 000. The RMSEs are averaged values over a period from t=25 to t=125 and over ten random number sequences. They are plotted against the covariance inflation factor for three covariance localization radii, which are indicated by two-digit numbers in the legends (unit: grid points). The minimum values of RMSE for each EnKF are shown by light blue markers.
Fig. 9. Comparison of RMSEs (solid lines) and spreads (dashed lines) of analysis of the PV-based EnKF using the first-order direct inversion (dark blue), the PV-based EnKF using the quasi-geostrophic inversion (light blue) and the conventional EnKF (light green) for ensemble sizes of (a) 25, (b) 50, (c) 100, and (d) 1 000. They are averaged values over ten random number sequences. The covariance inflation factor and localization radius are optimized for each ensemble size.
Fig. 9. (Continued)