A Constraint Method for Supersaturation in a Variational Data Assimilation System

KEN SAWADA^a AND YUKI HONDA^b

^a Meteorological Research Institute, Japan Meteorological Agency, Tsukuba, Ibaraki, Japan

^b Japan Meteorological Agency, Chiyoda-ku, Tokyo, Japan

(Manuscript received 30 October 2020, in final form 27 August 2021)

ABSTRACT: The reproducibility of precipitation in the early stages of forecasts, often called a spindown or spinup problem, has been a significant issue in numerical weather prediction. This problem is caused by moisture imbalance in the analysis data, and in the case of the Japan Meteorological Agency's (JMA's) mesoscale data assimilation system, JNoVA, we found that the imbalance stems from the existence of unrealistic supersaturated states in the minimal solution of the cost function in JNoVA. Based on the theory of constrained optimization problems, we implemented an exterior penalty function method for the mixing ratio within JNoVA to suppress unrealistic supersaturated states. The advantage of this method is the simplicity of its theory and implementation. The results of twin data assimilation cycle experiments conducted for the heavy rain event of July 2018 over Japan show that—with the new method—unrealistic supersaturated states are reduced successfully, negative temperature bias to the observations is alleviated, and a sharper distribution of the mixing ratio is obtained. These changes help to initiate the development of convection at the proper location and improve the fractions skill score (FSS) of precipitation in the early stages of the forecast. From these results, we conclude that the initial shock caused by moisture imbalance is mitigated by implementing the penalty function method, and the new moisture balance has a positive impact on the reproducibility of precipitation in the early stages of forecasts.

KEYWORDS: Atmosphere; Data assimilation; Mesoscale models; Numerical weather prediction/forecasting; Optimization

1. Introduction

The reproducibility of precipitation in the early stages of forecasts, especially for the spindown problem (spurious excessive precipitation), has been an issue in numerical weather prediction since the resolution of forecast models reached the convective scale, i.e., 1–10 km. Weak reproducibility is thought to be caused by moisture imbalance among the variables in the initial data (Hólm et al. 2002; Dixon et al. 2009). The imbalance in the initial data results from incompleteness of the assimilation procedure, which infers the plausible model state from observations and previous forecasts.

This issue is analogous to that of spurious gravity waves caused by the imbalance between the mass and flow fields in the initial data, the resolution of which is relatively large, i.e., 20–100 km (Daley 1991). For the mass–wind imbalance issue, by introducing a process called “initialization” before the forecast to damp the spurious gravity waves (Daley 1991), and by incorporating this initialization and the nonlinear mass balance related to hydrostatic equilibrium into the assimilation procedure (Courtier and Talagrand 1990; Gauthier and Thépaut 2001), this imbalance issue has almost been resolved.

However, the moisture imbalance issue remains unresolved. It is thought that the difficulty with moisture imbalance is due to the following: 1) proper moisture processes in the forecast model at the current convective resolution are unknown, 2)

simplified (or absent) moisture processes in the assimilation step do not have sufficient accuracy, and 3) an assimilation window of several hours is too long, considering the nonlinearity of the moisture processes (Stiller and Ballard 2009). In addition, the inhomogeneity and local variability of the moisture field make the problem difficult. The distribution of the relative humidity—which has upper and lower bounds, as well as a maximum near the upper bound in the lower and middle troposphere—requires a non-Gaussian error distribution of the moisture variable in the assimilation step. The assimilation procedure is usually constructed within the framework of a Gaussian error distribution, and while a background error covariance based on a Gaussian distribution is suitable for describing the smooth large-scale balance, which is well known and well modeled, it sometimes fails to describe the convective-scale balance.

The major method for handling the moisture imbalance problem is to construct proper control variables (parameters). This involves selecting valid moisture variables that are convenient for considering the balance with other variables at the convective scale (depending on the numerical models or the atmospheric states) and/or applying nonlinear transformations to obtain a Gaussian-like error distribution (Hólm et al. 2002; Gustafsson et al. 2018; Ingleby et al. 2013). To improve the quality of the analysis through effective use of satellite data, introducing the condensate water—in addition to water vapor—as a control variable in the moisture processes of the assimilation

Denotes content that is immediately available upon publication as open access.

Corresponding author: Ken Sawada, ksawada@mri-jma.go.jp

This article is licensed under a [Creative Commons Attribution 4.0 license](http://creativecommons.org/licenses/by/4.0/) (<http://creativecommons.org/licenses/by/4.0/>).

procedure has been attempted, and the effect has been discussed by [Gong and Hólm \(2011\)](#). Nudging techniques and incremental analysis update (IAU) are also useful tools for taking in moisture-related information ([Jones and Macpherson 1997](#); [Golding et al. 2014](#); [Bloom et al. 1996](#)). Recent progress and challenges in the assimilation of atmospheric water, including these methods, have been comprehensively reviewed by [Bannister et al. \(2019\)](#).

In the JMA mesoscale data analysis system called JNoVA ([Honda et al. 2005](#)), the existence of unrealistic supersaturated and negative moisture states in the minimal solution of the cost function (which differs from the analysis data in that it has lower resolution and limited atmospheric variables) is the main cause of moisture imbalance. Supersaturated states are not totally unrealistic in the atmosphere, but the supersaturation rarely exceeds 1% (i.e., the relative humidity is less than 101%) because of the existence of aerosols or condensation nuclei ([Rogers and Yau 1989](#)). In the minimal solution given by JNoVA, however, there are regions where the supersaturation sometimes exceeds 20% (see [section 3b](#)). Such states are considered to be unnatural, but they are inevitable in a general variational assimilation framework because of the features of the moisture distribution and the simplifications in the assimilation procedure mentioned above. Even in the simplest one-dimensional optimal interpolation (OI) data assimilation system for the relative humidity, supersaturated states can occur due to the correlation of the background error covariance matrix, if there is no way to take into account the local variability of the moisture-related variables. In addition, supersaturated states are not allowed in the cloud physics scheme of the JNoVA forecast model—supersaturated moisture is converted into rain or cloud water and latent heat is released at each time step of the evolution (saturation adjustment). In practice, from the minimal solution, the analysis data are obtained by increasing the resolution and supplementing with unanalyzed atmospheric variables from the first guess. Nonphysical supersaturated and negative moisture states in the minimal solution are inherited to the analysis data, but they are removed by preprocessing when the initial data are made from the analysis data. This preprocessing is intended to avoid forecast crashes due to numerical instability caused by the excessive moisture. However, preprocessing cannot take moisture balance into account, and it sometimes causes another moisture imbalance that results in a spinup problem—underestimation of precipitation in the early stages of the forecast—rather than a spindown problem. Thus, in this paper we aim to obtain a minimal solution that has a more realistic moisture balance and that suits our forecast model by suppressing the unrealistic moisture states better in the minimizing process of the variational method. For this purpose, we propose implementing an exterior penalty function method in the constrained optimization problem. This method is similar to that discussed in [Courtier and Talagrand \(1990\)](#) and [Gauthier and Thépaut \(2001\)](#), which uses a weak constraint in order to reduce the mass/wind imbalance, but this penalty function method treats the boundedness of the moisture-related variable directly as a strict constraint.

TABLE 1. Observations used in JNoVA.

	Kind	Element
In situ obs	Land surface obs	Pressure, precipitation
	Automated weather station	Precipitation
	Sea surface obs	Pressure
	Upper-air sounding	Pressure, temperature, wind, relative humidity
	Aircraft	Temperature, wind
Remote sensing	Wind profiler	Wind
	Radar	Precipitation, wind
	Ground-based GPS	Integrated precipitable water
Satellite	Advanced imager	Atmospheric motion vector
	Microwave sounder	Brightness temperature
	Microwave imager	Precipitation
	Scatterometer	Sea surface wind

The rest of this paper is organized as follows. In [section 2](#), the minimization process in the assimilation procedure is reformulated as a constrained optimization problem, and the exterior penalty function method is introduced. The results of twin experiments to examine this procedure are presented in [section 3](#), and our conclusions are given in [section 4](#).

2. Penalty function method

The analysis is performed statistically by using an assimilation procedure based on the background and the observation error covariances. Thus, it is necessary to provide proper error covariances, which can describe the appropriate moisture balance at the convective scale (with a proper choice of control variables), in the assimilation procedure. However, this is extremely difficult, as pointed out by [Bannister et al. \(2019\)](#). In JNoVA the main cause of moisture imbalance in the analysis data is the existence of supersaturated and negative moisture states in the minimal solution, so a method that can avoid generating such unrealistic moisture states in the assimilation procedure without breaking the moisture balance or entering into the problem of reconfiguring the error covariances is very attractive. This can be done by confining the minimal solution to the subset of the control variable space where the relative humidity is in the range from 0% to 100% in the minimization process within the variational assimilation procedure.

This kind of problem can be formulated as a constrained optimization problem:

$$\min_{\mathbf{x} \in X} f(\mathbf{x}) \quad \text{subject to } g(\mathbf{x}) \leq 0, \quad (1)$$

where $f(\mathbf{x})$ is the objective function (or cost function), $g(\mathbf{x})$ is the (nonlinear) constraint function, and X is the control space. The region $F = \{\mathbf{x} \in X | g(\mathbf{x}) \leq 0\}$ is called the feasible region, and the region $\{\mathbf{x} \in X | g(\mathbf{x}) = 0\}$ is the effective region. To deal with this problem mathematically, we have to check a suitable constraint qualification. This is very hard for our present problem, so we assume that the functions f and g have desirable properties such as continuity, differentiability, and convexity in order to proceed.

FIG. 1. The behaviors of the cost values as functions of the number of iterations in the minimization process, with enlarged views shown in the insets in each panel. (a) Total costs of Ctrl and the tests. (b) Costs of the background, observation, and penalty terms in the cost function in (5). (c) Costs of the temperature and relative humidity elements in the observation term and of the penalty for the mixing ratio.

Numerous algorithms have been developed to solve the problem (1); see Bertsekas (1982); Bazaraa et al. (2006). Among them, the exterior penalty function method is easy to implement within the existing minimization process in the variational assimilation procedure. In this method, the original constrained problem, (1), is converted to an unconstrained problem,

$$\min_{\mathbf{x} \in X} f_2(\mathbf{x}), \quad (2)$$

by introducing the auxiliary function defined by

$$f_2(\mathbf{x}) = f(\mathbf{x}) + \lambda \max\{0, g(\mathbf{x})\}^\alpha, \quad (3)$$

where $\lambda > 0$ and $\alpha \geq 1$ are the penalty parameters. The second term of Eq. (3) is sometimes called the “penalty term” or “penalty function.” If there is more than one constraint, one additional penalty term is added for each constraint.

The concept of the exterior penalty function method is to add some penalty value if $\mathbf{x}$ is outside the feasible region F but not if $\mathbf{x}$ is inside F . The penalty parameter α controls the effect of putting the solution back into the feasible region, depending on the degree of violation $\max\{0, g(\mathbf{x})\}$. It is popular to set $\alpha = 2$ to maintain continuity at the effective region, but in this case we have to solve a sequential minimization problem repeatedly while increasing λ to infinity, since the effect of the penalty term reduces quadratically as the state approaches the feasible region. A penalty function that yields the same solution as the original problem without altering the value of the parameter λ is called an exact penalty function. When $\alpha = 1$, we can get an exact penalty function by choosing a sufficiently large value for λ . In practice, a suitable value for λ depends on the gradients of f and g , and it is known that too large a value of λ sometimes destroys a minimization process based on gradient methods.

In our problem, there is arbitrariness in constructing the constraint function g on the moisture-related variable. Here we set

$$\begin{cases} g_{1i}(\mathbf{x}) = qv_i(\mathbf{x}) - qvs_i(\mathbf{x}) \\ g_{2i}(\mathbf{x}) = -qv_i(\mathbf{x}), \end{cases} \quad (4)$$

where qv_i and qvs_i are the water vapor mixing ratio (QV) and the saturation mixing ratio (QVS), respectively, at the i th grid point. Here, $g_{1i} \leq 0$ is the constraint for supersaturation at the i th grid point, and $g_{2i} \leq 0$ is that for the negative mixing ratio. Since the mixing ratio has large values in the lower troposphere, this construction is intended for the effective modification of the moisture field in the lower troposphere, which is closely related to the initiation and development of deep convection and the generation of precipitation.

3. Twin experiments

a. System

We employed an experimental system based on the JMA mesoscale forecast system to examine the effects of the penalty function method. This system is composed of a forecast component that employs a nonhydrostatic model called the mesoscale model (MSM), which is based on the Japan Meteorological Agency Nonhydrostatic Model (JMA-NHM:

FIG. 2. Vertical profile of violation. The bold lines in the positive sector of each panel are for supersaturated states, and the thin lines in the negative sector are for negative moisture states. (a) The violation rate defined as the ratio of the number of violation points to the total number of horizontal grid points. (b) (left) Horizontal summation of the violations and (right) an enlarged view.

Saito et al. 2006, 2007; Saito 2012) and an assimilation component that contains the four-dimensional variational data assimilation system called JNoVA (Honda et al. 2005; Gustafsson et al. 2018), which was operational until March 2020. The MSM has 817×661 horizontal grid points with 5-km resolution and 76 vertical layers extending up to 21.8 km in height, and it covers Japan and the surrounding area. The JNoVA component adopts an incremental approach (Courtier et al. 1994); its inner loop model has 271×221 horizontal grid points with 15-km resolution and 38 vertical layers covering the same region as the MSM. The assimilation window is 3 h, and the control variables are the eastward wind, northward wind, surface pressure, potential temperature, and a pseudo relative humidity defined by scaling the mixing ratio by the background saturation mixing ratio. Background error cross correlations between the moisture variable and the other variables are not considered. Hydrometeors are not included among the control variables, and perturbations of the hydrometeors are ignored in the inner forward model. The list of observations used in JNoVA is shown in Table 1. We have many observations related to the temperature field but not so many observations related to the moisture field, which were mainly obtained by radiosonde observations.

FIG. 3. (a) Horizontal distribution of the relative humidity (%), corresponding to the scale represented by the color bar at the bottom, for (top left) Ctrl and (middle),(bottom) the tests (L100, L200, L500, and L1000) at the model level $kz = 14$ (about 2500 m in height). (b) Vertical relative humidity profile at 33.5°N, 128.5°E, which is near the center of the high relative humidity region above the sea west of Kyushu (indicated by the black arrow in the panel for Ctrl).

The inner forward model of JNoVA is a simplified nonlinear JMA-NHM with a large-scale condensation scheme and the Kain–Fritsch (KF) convective scheme (the KF scheme is not considered in the adjoint step). The outer model of JNoVA, which has 5-km resolution, is the full JMA-NHM—with a cloud-physics scheme based on a 6-class 3-ice bulk scheme, and a modified KF scheme—which is the same as in the MSM. Because of differences in the moist physics, the forecasts of the inner and outer models differ from each other, but there is not as much difference between the two as the difference caused by the existence of supersaturated states. This has been checked by conducting experiments with and without preprocessing, as discussed in [section 3c](#). For this reason, we do not consider further the effects that result from differences in the moist physics.

Supersaturated states in the analysis data are converted into just 100% relative humidity states, and negative moisture

states are converted into zero-moisture states by preprocessing before running the outer model to generate the initial conditions for the MSM. Preprocessing is employed to avoid forecast crashes due to numerical instability. However, preprocessing merely cuts (or adds) extra (or missing) moisture, so it may cause another moisture imbalance.

The implementation of the exterior penalty function method in the variational assimilation system is very simple, and the auxiliary function—or the cost function in the unconstrained form—is given by

$$\begin{aligned}
 J(\mathbf{x}_0) = & \frac{1}{2}(\mathbf{x}_0 - \mathbf{x}_0^b)^T \mathbf{B}^{-1}(\mathbf{x}_0 - \mathbf{x}_0^b) \\
 & + \frac{1}{2} \sum_n [\mathbf{y}_n^o - H_n(M_{0 \rightarrow n} \mathbf{x}_0)]^T \mathbf{R}^{-1} [\mathbf{y}_n^o - H_n(M_{0 \rightarrow n} \mathbf{x}_0)] \\
 & + J_{\text{dri}}(\mathbf{x}_0) + J_{\text{qv}}(\mathbf{x}_0)
 \end{aligned} \quad (5)$$

FIG. 4. The difference between Ctrl and L200. (a) (left) The difference in the mixing ratio [$L200 - Ctrl$ ($kg\ kg^{-1}$); shaded as in the color bar at the bottom] and contour lines of the mixing ratio of L200 (black lines) at model level $kz = 14$ (about 2500 m in height), and (right) a vertical cross section along the longitude $128.5^\circ E$. The regions with increased mixing ratio are indicated by the black arrows. (b) As in (a), but for the temperature (K). The regions surrounded by white and yellow lines show the boundaries of the supersaturated regions of Ctrl and L200, respectively.

in JNoVA, where

$$J_{qv}(\mathbf{x}_0) = \lambda \sum_i [\max\{0, g_{1i}(\mathbf{x}_0), g_{2i}(\mathbf{x}_0)\}]^\alpha. \quad (6)$$

Here, $\mathbf{x}_0^b$ is the background state at the initial time; $\mathbf{y}_n^o$ is the observation at the n th time step; $\mathbf{B}$ and $\mathbf{R}$ are the forecast and observation error covariance matrices, respectively; H_n is the observation operator; $M_{0 \rightarrow n}$ is the time development operator; and J_{dfi} is the penalty term for the incremental digital filter initialization (DFI) used to control spurious gravity waves. Actually, because JNoVA adopts an incremental approach, the equations above need slight changes, but we leave them unchanged for simplicity. In the forward step of JNoVA, the arguments of the penalty function QV and QVS [see Eqs. (4) and (6)] are obtained by recalculating the basic fields of the pressure and temperature right after updating the basic fields of control variables in each iteration of the minimization loop. In the adjoint step, the corresponding adjoint variables are calculated by dividing them into cases (supersaturated, negative moisture, otherwise) that are determined by the states of the basic fields.

The constraint defined by Eq. (4) has the effect of reducing the mixing ratio and increasing the saturation mixing ratio for

the supersaturated state at each grid point. Since the saturation mixing ratio is a monotonically increasing function of the temperature, this constraint has the effect of increasing the temperature. These effects are spread and developed through the minimization process by the covariance matrices $\mathbf{B}$ and $\mathbf{R}$ and by the nonlinear inner model, and they produce a new moisture balance in the minimal solution (and then the analysis data).

To investigate the impact of the penalty function method on the analysis and forecast, we conducted twin data assimilation cycle experiments. In the following, the experiment that employs the usual JNoVA system is called “Ctrl,” and those utilizing the new JNoVA system equipped with the penalty term J_{qv} are called “tests” (the penalty term J_{dfi} is included in both Ctrl and tests). In the tests, we set $\alpha = 1$ and $\lambda = 100, 200, 500$, and 1000, which are labeled “L100,” “L200,” and so on, in order to determine the effects of the parameter λ . We use preprocessing in both Ctrl and tests in order to remove the supersaturated or negative moisture states completely and to align the conditions when we make the initial data from the analysis data (see section 3b). The twin experiments extend from 28 June to 8 July 2018, which coincides with the period of the “heavy rain event of July 2018” that caused tremendous

FIG. 5. The 3-h accumulated precipitation [mm (3 h)^{-1}]; shaded as in the color bar] at 1200 UTC 28 Jun 2018. The different panels show (top left) the radar/rain gauge analyzed precipitation (RA), (top right) the first guess (Ges), and (remaining rows) the forecasts from Ctrl and the tests (L100, L200, L500, and L1000). In the RA panel, the region for which analyzed precipitation data are not available is shaded gray. The forecast from the Ctrl analysis data but without the preprocessing to remove supersaturated states is labeled “Ctrl_nochkqv.” We focus on the reproducibility of precipitation indicated by the black oval in the RA panel.

disasters in Japan. Clearly, it would be very useful to prevent or mitigate such disasters, which might be possible if we could catch the initiation of such severe events earlier.

b. Analysis

Here we discuss the impacts of the penalty function method on the assimilation procedure and the analysis data at the first cycle (0900 UTC 28 June). The behaviors of the cost values in the minimization process are shown in Fig. 1. The behavior of

the total cost seems not to be greatly affected by the new, added penalty term, but the convergence becomes slightly slower as the value of λ increases. Since the global minimum of the constrained problem is never less than that of the original problem with no constraint, the fact that the final total costs of L100 and L200 are smaller than that of Ctrl implies either that each final state may be a local minimum of the nonlinear, nonconvex cost function or that the number of iterations may not be sufficient. However, the final total costs of Ctrl and tests

FIG. 6. (a) Time evolution of the temperature increment $[(K)$; shaded as in the color bar at the bottom] of Ctrl at 700 hPa and contour lines of temperature (red lines). (b) As in (a), but for L200. FT indicates the forecast period (h). The large negative increment in Ctrl at FT = 0 is indicated by the black arrow.

are not very different—with the differences lying within the range of 0.2%—so we think that the effect of the imperfectness of the minimization process is quite small.

Figure 1b shows that most of the total cost comes from the observation term and that, compared with Ctrl, the tests are slightly closer to the background than to the observations in the minimization process. The values of the penalty terms for DFI

and QV are considerably smaller than the other terms. Figure 1c shows the costs of several elements, i.e., the temperature, the relative humidity, and the penalty for the non-physical mixing ratio (J_{qv}). In the first iteration (it = 0), the value of J_{qv} is always 0, since the mixing ratio q_i is within the range $[0, q_{vs_i}]$ at every grid point in the first guess. The values of J_{qv} , which are defined by the violation multiplied by

FIG. 7. As in Fig. 6, but for the mixing ratio (kg kg^{-1}) at 850 hPa. The positive increment near the sea west of Kyushu in Ctrl at FT = 0 is indicated by the black oval.

individual values of λ , are effectively suppressed, and they finally reach values around $O(10)$, regardless of the value of λ . The supersaturated states are more strongly suppressed in the tests that have large λ . By definition, the costs of the temperature and the relative humidity are thought to be strongly related to the constraint in the tests, but there are no obvious differences, even though L200 shows a relatively good fit to the observations. This is because the costs shown in Fig. 1 are

evaluated over the whole domain of the MSM, while the region affected by the penalty method is only a part of the MSM domain, as discussed later.

Too large a value of λ may destroy the minimization process based on the gradient method, but in the range $\lambda \leq 1000$, the minimization process seems to work well. In an additional test with $\lambda = 10\,000$, the behavior of the cost is very different from those of Ctrl and the other tests, and the final cost remains large

FIG. 8. The 1-h accumulated precipitation [(mm h^{-1}) ; shaded as in the color bar] (left to right) from 1000 UTC (FT = 1) to 1200 UTC (FT = 3) 28 Jun 2018. (a) RA, (b) Ctrl, and (c) L200. The line segment AB is the location of the cross section shown in Fig. 9.

(not shown). Thus, although it is difficult to determine the limit of λ , at least we can say that 10 000 is too large for our system.

Figure 2 shows the impact of the penalty function J_{qv} on the mixing ratio modification in the minimal solutions. The violation points—which are in supersaturated or negative moisture states and have nonzero J_{qv} —are substantially reduced by the penalty function method as the value λ becomes large. While the violation ratio for Ctrl exceeds 5% in the lower troposphere, those of the tests are less than 0.2%. In the upper troposphere and lower stratosphere, more than 1% of the violation points remain in the tests, in agreement with the discussion about the construction of the constraint function. Figure 2b shows the horizontal summations of the violations, given by $\sum_i \max\{0, g_{1i}(\mathbf{x}_0), g_{2i}(\mathbf{x}_0)\}$, as a measure of the net impact. The violation is quite effectively suppressed in all the tests, even in L100. In addition, we can see that both the violation rate and the horizontal summation of the violations for negative moisture states are much smaller than those for supersaturated states. Consequently, we hereafter ignore the effect caused by negative moisture.

The relative humidity distributions in the minimal solutions of Ctrl and the tests presented in Fig. 3 show that the spurious supersaturated regions in Ctrl—sometimes exceeding the

relative humidity of 120%—are removed effectively in the tests, especially in those with larger values of λ . Of course, the supersaturated regions are not completely eliminated over the whole MSM area even in L1000, as indicated in Fig. 2, and there are slight differences in the size or location of the supersaturated region among the tests. However, the distributions of relative humidity in the tests are very similar to each other on the whole; in particular, it is hard to find a difference between L500 and L1000. This may mean that the values of the parameter $\lambda = 500$ and 1000 are large enough for this atmospheric field.

Note that modifications of the atmospheric fields due to the use of the penalty function method do not occur only in the supersaturated regions of the tests shown in Fig. 3, since those regions occur just in the final state of the minimization process. The constraints affect the atmospheric fields, while changing the sizes and locations of the supersaturated regions during the minimization process. The supersaturated regions of the tests are actually generated and spread near the centers of the supersaturated regions of Ctrl in the early stages of the minimization process and then dissipate gradually. The peaks of the penalty term J_{qv} in Fig. 1c are affected by this behavior.

Figure 4 shows the effects of the constraint on the atmospheric fields in the minimal solution. Compared with Ctrl, the

FIG. 9. Vertical p velocity [hPa h^{-1}]; shaded as in the color bar] on the vertical cross section along the line segment AB in Fig. 8. (a) Ctrl and (b) L200, showing contour lines of the equivalent potential temperature (red lines) and the in-plane circulating wind (black arrows).

effects of reducing the mixing ratio and increasing the temperature in the L200 tend to become clearer near the centers of the supersaturated regions of Ctrl, rather than in the analogous regions of the L200. These effects coincide with that expected from the construction of the constraint function and from previous discussions. We also note that there are small but clear regions with increased mixing ratio as shown in the right panel of Fig. 4a, without exceeding the saturation level due to the increased temperature. These features of the atmospheric

fields in the minimal solution are inherited to the analysis data. Analyzing a sharper distribution with increased mixing ratio in appropriate locations of the lower-level troposphere may have a significant impact on the initiation and development of deep convection and on the generation of precipitation.

c. Forecast

Here we consider the effect of the penalty function method on forecasts at the first cycle. Figure 5 shows the 3-h accumulated

FIG. 10. (a) (left) Profile of cycle-mean horizontal summation of the violation and (right) an enlarged view. The bold lines in the positive sector are for supersaturated states, and the thin lines in the negative sector are for negative moisture states. (b) The cycle-mean difference between Ctrl and the tests (tests – Ctrl) for (left) the mixing ratio and (right) the temperature.

precipitation in the twin experiments from the initial time of 0900 UTC 28 June 2018. In this event, we see clear, strong rainfall along the baiu front in the sea northwest of Kyushu, in the radar/rain gauge analyzed precipitation data (RA, treated as the observations). This strong rainfall is not seen in the first guess, and it is not adequately reproduced in the forecast from Ctrl. We also see that weak reproducibility in Ctrl is caused to some extent by the above mentioned preprocessing. In the forecast from the tests, the reproducibility (including the location, distribution, and amount) of precipitation is improved. The distributions of precipitation in the tests (L200, L500, and L1000) are quite similar to each other, and the maximum amounts of the precipitation in

L500 and L1000 are almost comparable to that of the RA. Although it seems that in this case L500 and L1000 reproduce the precipitation better than L100 or L200, the differences from Ctrl are all quite big. Consequently, we focus conservatively hereafter on the test of L200 (we consider the effect of the value of λ later).

In the previous discussion, we have checked the differences between Ctrl and the tests in the analysis. However, the differences—the warmer temperatures and lower mixing ratios around the supersaturated region shown in Fig. 4—cannot directly explain the differences in precipitation shown in Fig. 5. Thus, we have to investigate the time evolution of the atmospheric fields in Ctrl and the tests carefully.

FIG. 11. (a) Evolution of the cycle-mean increments of Ctrl (lines) and of L200 (dashed lines) in the assimilation window. The increments in (left) the mixing ratio (kg kg^{-1}) and (right) the temperature (K) at every 1-h forecast period are shown. (b) The 1-h trends in the forecast of (left) the mixing ratio ($\text{kg kg}^{-1} \text{h}^{-1}$) and (right) the temperature (K h^{-1}) for Ctrl (lines) and for L200 (dashed lines) in the assimilation window.

Figure 6 shows the time evolution of the temperature increment at 700 hPa. There is a large, obvious, negative increment at the initial time in Ctrl (indicated by the black arrow), while the corresponding increment in L200 is more moderate. The strong, cool increment of Ctrl vanishes rapidly in the time evolution from FT = 0 to FT = 1 (where FT indicates the forecast range in hours), and the time evolution behaviors of Ctrl and L200 are relatively similar to each other in the range from FT = 1 to FT = 3. Thus it is possible that the heating effect produced by the constraint may help to mitigate the initial imbalance.

The time evolution of the mixing ratio increment at 850 hPa is shown in Fig. 7. Note that the supersaturated states are

removed by the preprocessing, so the differences between Ctrl and L200 at the initial time of the forecast are smaller than those of the analysis data. In Ctrl, the initial positive increment near the sea west of Kyushu (indicated by the black oval) is rapidly reduced by FT = 1, and it moves along with the southwest flow. On the other hand, in L200 the corresponding initial positive increment moves along with the flow without changing rapidly, and advection of air with a high mixing ratio from the southwest in the sea west of Kyushu results in the strong precipitation at FT = 2 and 3 shown in Fig. 5. This high mixing ratio region is supported by advection from the lower layers, as discussed later. In addition, compared with Ctrl, the high mixing ratio region of L200 in the sea west of Kyushu

seems to trigger the convection. These features in the mixing ratio also imply that L200 reduces the initial imbalance and gathers the moisture to organize precipitation at the proper location.

The evolution of the 1-h accumulated precipitation is shown in Fig. 8. In Ctrl, although the region of weak precipitation is widely distributed at $FT = 1$, heavy precipitation like that seen in RA does not appear until $FT = 3$. In L200, there also is no heavy precipitation at $FT = 1$, and the precipitation area of L200 is smaller than that of Ctrl. However, the precipitation begins to strengthen at $FT = 2$, and heavy precipitation occurs at $FT = 3$, the location of which is very close to that of the RA. The organization of the precipitation can be seen in Fig. 9, which shows a vertical cross section along the line segment AB in Fig. 8. Although there is no area with clear upward vertical velocities in Ctrl, L200 displays an area of apparent upward vertical velocity that is supported by the advection of air with a high equivalent potential temperature from the lower troposphere, which forms the convection related to the precipitation.

d. Cycle experiments

Here, we examine the effect of our new method using cycle experiments. We conducted cycle experiments for Ctrl and for the tests (L100, L200, L500, L1000) from 1200 UTC 28 June to 1200 UTC 8 July. The analyses were performed every 3 h (we obtained 81 analyses), and then we performed 12-h forecasts from these analyses every 12 h (21 forecasts). Figure 10a shows that the cycle-mean values of the violation in the tests are effectively reduced compared to Ctrl, as in the case of the first cycle analysis shown in Fig. 2b. Figure 10b shows that the cycle-mean value of the mixing ratio in the minimal solutions is reduced mainly in the middle level of the troposphere by the new method, while the cycle-mean temperature is increased mainly in the lower troposphere. These are consistent with the theory described in section 2 and the results in the first cycle. In addition, the profiles of the tests shown in Fig. 10b are similar to each other, and the differences among the tests are not very large compared to the daily variations of the mixing ratio or temperature. This implies that effect of the constraint is introduced adequately even in L100 through cycle assimilation and that the differences in the remaining violation are sufficiently small.

The evolution of the cycle-mean increments and of the 1-h forecast trends of the mixing ratio and the temperature for Ctrl and L200 are shown in Fig. 11. The positive mean increments for the mixing ratio at the beginning of the assimilation window ($FT = 0$) are reduced in L200, and the profile of the increment for L200 at $FT = 0$ approaches the profiles at the other times, while the negative trend of Ctrl at $FT = 0$ is mitigated in L200, except near the surface. For the temperature, there are relatively large positive increments in the lower troposphere (below 850 hPa) at $FT = 0$ compared to Ctrl, but the trend of L200 at $FT = 0$ is smaller than that of Ctrl except near the surface and the upper troposphere. These behaviors imply that the initial shock is reduced in L200 mainly in the lower and middle troposphere, and that the new method provides an analysis data that is consistent with the forecast model. In addition,

FIG. 12. Fractions skill scores (FSSs) of the forecasts up to 12 h. The verification grid is 10 km, and the precipitation thresholds are (a) 1.0 and (b) 10 mm h^{-1} . The forecast range 0 indicates the time at end of the assimilation window.

there is a possibility of improving the first guess through cycle assimilation.

We also performed verifications of the 12-h precipitation forecasts from Ctrl and the tests. The forecasts are initiated from the end of the assimilation window every 12 h (0000 and 1200 UTC), during the cycle period. The fractions skill scores (FSSs) with a 10-km verification grid and thresholds of 1.0 and 10.0 $mm h^{-1}$ are shown in Fig. 12. We can see clear improvements in the FSSs of the tests at both thresholds in the early stages of the forecast. For other cases with different verification grid sizes and thresholds, we confirmed that in general the tests are superior to Ctrl. Although L1000 seems to be better among the tests from Fig. 12, the scores vary with the threshold and the atmospheric state, and the dependence of the FSS on the value of λ is not straightforward. Therefore, we cannot determine what value of λ gives the best improvement in the FSS, but we can say that the improvement is robust for the values of λ we considered. One possibility is that the value $\lambda = 100$ may be sufficient in the case of cycle assimilation and the differences among the tests are small compared to the differences between Ctrl and tests. This is consistent with the discussion of the results shown in Fig. 10.

FIG. 13. (a) (top) Horizontal and (bottom) vertical distributions of the observed temperatures, together with the residual errors [observation – analysis data: $O - A$ (K); colored according to the color bar at right] from Ctrl, which are used in the assimilation. The radiosonde data are marked with black circles. (b) Histograms of the residual temperature errors from (left) Ctrl and (right) L200. The white vertical bars represent the observations in the region 30° – 37° N, 123° – 136° E; that is, for the whole region shown in (a). The gray bars represent the observations in the region covering 32° – 35° N, 126° – 131° E, and 300–950 hPa, which is indicated by the gray box in (a). The quantities “me” and “sd” are the mean and standard deviation of the data, respectively.

FIG. 14. Vertical profiles of the residual temperature errors [observation – analysis data: $O - A$ (K)] compared to the radiosonde data for (a) Fukuoka, (b) Matsue, and (c) Kagoshima.

e. Validation

The penalty function method appears to bring a new moisture balance into the analysis, providing a positive impact on the reproducibility of the precipitation. However, the change in the temperature field is not small, so it is necessary to check whether this change is consistent with the observations.

Figure 13a shows the location of the temperature observations obtained from 0900 to 1200 UTC 28 June, which are used in the data assimilation at the first cycle in Ctrl and the tests. Note that these data are mostly airborne measurements and radiosonde observations, and the number of observations in the supersaturated region of Ctrl shown in Fig. 3 is quite small. Figure 13b shows histograms of the residual temperature errors (observation – analysis data: $O - A$) over the whole region shown in Fig. 13a and for the limited space around the supersaturated region outlined by the gray box in Fig. 13a. The temperature from the Ctrl analysis data tends to be cooler than the observations, and the deviation (bias) is large around the supersaturated region. The deviations in both these regions are slightly reduced in L200, but there is no significant difference in the modifications between the two regions. Because of the evolution of the atmospheric field, the effects of the penalty function J_{qv} spread, as shown in Figs. 6 and 7, and the modification does not appear to be limited solely to the supersaturated region at the initial time.

Figure 14 shows the residual temperature errors compared to the radiosonde temperatures observed at 1200 UTC 28 June. Examination of the residual errors at the Fukuoka radiosonde site shows that the low temperature bias in the 400–600-hPa range that occurs in the first guess and in the Ctrl analysis data are reduced in the tests. The vertical profiles of the tests are similar to each other, but the differences among them are large around 700 hPa. Because this comparison is done at the end of the assimilation window, it is thought that those differences may reflect a slight misalignment of the high temperature region due to the time development.

In the comparison, at the Matsue radiosonde site the differences from Ctrl occur mainly between 500 and 300 hPa, and the profiles of the tests are much more similar to each other.

At this location, the lower-level temperature differences due to the supersaturated region at the initial time flow away, and air that originates from the nonsupersaturated region comes in. In the comparison at Kagoshima, there are no large differences among the tests or even with Ctrl, since the location of the radiosonde station is far from the supersaturated region.

These results indicate that the effect of the penalty function method is mainly limited to the vicinity of the supersaturated region and the region downstream from it, as expected theoretically, and that this method reduces the residual errors statistically. However, there may sometimes be disagreements with the radiosonde data because of local variations in the atmospheric fields.

Figure 15 shows the profiles of the means and the root-mean-square errors of innovation (observation – first guess: $O - B$) for the temperature and the relative humidity, during the cycle assimilation period. Except near the surface, the tests in general show a better fit to the observations for both the temperature and the relative humidity. We also see that the differences among the tests are small, especially for the temperature. These imply that the changes in the temperature and the relative humidity field in the tests are valid and that the effect is robust for the values of λ we considered.

4. Summary

We have introduced an exterior penalty function method into the minimization process employed in variational assimilation in order to suppress nonphysical supersaturated or negative moisture states and to reduce the moisture imbalance in the analysis data that may cause a degradation of the precipitation reproducibility. The exterior penalty function method is a numerical algorithm used for solving constrained optimization problems, and the simplicity of implementing it within the assimilation procedure is a significant advantage. However, to save computational costs we have to choose a proper value of the parameter λ —taking into account the sizes of the gradients of other terms in the cost function—when we use this method as an exact penalty function method.

FIG. 15. The (left) mean (ME) and (right) root-mean-square error (RMSE) of innovation (observation – first guess: $O - B$) of (a) temperature (K) and (b) relative humidity (%) during the cycle experiment period.

The results of twin experiments for the heavy rain event of July 2018, show that the penalty function method effectively suppresses nonphysical supersaturated states in the analysis data—with changes supported mainly by increasing the temperature and reducing the mixing ratio around the supersaturated region compared with Ctrl—as expected theoretically. The forecast initiated from the analysis data by this method shows that the initial shock in the temperature and mixing ratio is mitigated and produces convection at the proper location, and it also gives improvement in the precipitation FSS. These improvements indicate that the penalty function method provides a preferable analysis data, with new moisture balance mainly affecting the mixing ratio and temperature near the supersaturation area.

Comparison with the observations shows that this method reduces the departures ($O - B$ and $O - A$) statistically, and gives better agreement with the observed data, except near the surface.

The choice of the value of the parameter λ is directly related to the violation of the constraint. In the first cycle, the value $\lambda = 100$ seemed inadequate, because the supersaturated region remained to some extent and the improvement in precipitation reproducibility is small. However, during the cycle period, the differences from different λ values are relatively small. We speculate that the effect of the constraint is introduced

adequately even in the case $\lambda = 100$ through cycle assimilation, and the differences in the remaining violations or atmospheric states are small enough compared to the effects caused by other incompleteness in the assimilation procedure. We also emphasize that the reproducibility of the precipitation depends strongly on schemes or parameterizations in the physical processes of the forecast model. These are adjusted to the analysis data of Ctrl, and some kind of adjustment also may be needed to obtain the benefit of the penalty function method more effectively.

We have investigated the effects of the penalty function method using a cycle assimilation for a specific event and have demonstrated its contribution to reducing the moisture imbalance. In future, we plan to investigate the effects on the analysis of and forecasts for other events under various atmospheric conditions, taking into consideration the relevance to the error distributions of the moisture-related variables statistically. We also plan to examine the effects of the new method on the new JMA operational mesoscale assimilation system called “asuca-Var,” which is based on a more advanced non-hydrostatic forecast model than JMA-NHM and that includes various changes and improvements in the assimilation procedure.

Acknowledgments. The authors are grateful to Hiromu Seko, Takuya Kawabata, Tadashi Fujita, Yasutaka Ikuta, and Daisuke Hotta in the Meteorological Research Institute for helpful discussion. This work was supported by Japan Society for Promotion of Science (JSPS) Grant JP 19K23468 and by MEXT as “Program for Promoting Researches on the Supercomputer Fugaku” (Large Ensemble Atmospheric and Environmental Prediction for Disaster Prevention and Mitigation, hp200128).

REFERENCES

- Bannister, R. N., H. G. Chipilski, and O. Martinez-Alvarado, 2019: Techniques and challenges in the assimilation of atmospheric water observations for numerical weather prediction towards convective scales. *Quart. J. Roy. Meteor. Soc.*, **146**, 1–48, <https://doi.org/10.1002/qj.3652>.
- Bazaraa, M. S., H. D. Sherali, and C. M. Shetty, 2006: *Nonlinear Programming: Theory and Algorithms*. 3rd ed. John Wiley and Sons, 872 pp.
- Bertsekas, D. P., 1982: *Constrained Optimization and Lagrange Multiplier Method*. Academic Press, 395 pp.
- Bloom, S., L. Takacs, A. D. Silva, and D. Ledvina, 1996: Data assimilation using incremental analysis updates. *Mon. Wea. Rev.*, **124**, 1256–1271, [https://doi.org/10.1175/1520-0493\(1996\)124<1256:DAUIAU>2.0.CO;2](https://doi.org/10.1175/1520-0493(1996)124<1256:DAUIAU>2.0.CO;2).
- Courtier, P., and O. Talagrand, 1990: Variational assimilation of meteorological observations with the direct and adjoint shallow-water equations. *Tellus*, **42A**, 531–549, <https://doi.org/10.3402/tellusa.v42i5.11896>.
- , J. N. Thépaut, and A. Hollingsworth, 1994: A strategy for operational implementation of 4D-Var, using an incremental approach. *Quart. J. Roy. Meteor. Soc.*, **120**, 1367–1387, <https://doi.org/10.1002/qj.49712051912>.
- Daley, R., 1991: *Atmospheric Data Analysis*. Cambridge University Press, 457 pp.
- Dixon, M., Z. Li, H. Lean, N. Roberts, and S. Ballard, 2009: Impact of data assimilation on forecasting convection over the United Kingdom using a high-resolution version of the Met Office

- Unified Model. *Mon. Wea. Rev.*, **137**, 1562–1584, <https://doi.org/10.1175/2008MWR2561.1>.
- Gauthier, P., and J.-N. Thépaut, 2001: Impact of the digital filter as a weak constraint in the preoperational 4DVAR assimilation system of Météo-France. *Mon. Wea. Rev.*, **129**, 2089–2102, [https://doi.org/10.1175/1520-0493\(2001\)129<2089:IOTDFA>2.0.CO;2](https://doi.org/10.1175/1520-0493(2001)129<2089:IOTDFA>2.0.CO;2).
- Golding, B. W., and Coauthors, 2014: Forecasting capabilities for the London 2012 Olympics. *Bull. Amer. Meteor. Soc.*, **95**, 883–896, <https://doi.org/10.1175/BAMS-D-13-00102.1>.
- Gong, J., and E. Hölm, 2011: Development of cloud condensate background errors. *ECMWF Newsletter*, No. 128, ECMWF, Reading, United Kingdom, 23–27.
- Gustafsson, N., and Coauthors, 2018: Survey of data assimilation methods for convective-scale numerical weather prediction at operational centers. *Quart. J. Roy. Meteor. Soc.*, **144**, 1218–1256, <https://doi.org/10.1002/qj.3179>.
- Hölm, E., E. Andersson, A. Beljaars, P. Lopez, J.-F. Mahfouf, A. Simmons, and J.-N. Thépaut, 2002: Assimilation and modelling of the hydrological cycle: ECMWF's status and plans. ECMWF Tech. Memo. 383, <https://doi.org/10.21957/kry8prwuq>.
- Honda, Y., M. Nishijima, K. Koizumi, Y. Ohta, K. Tamiya, T. Kawabata, and T. Tsuyuki, 2005: A pre-operational variational data assimilation system for a non-hydrostatic model at the Japan Meteorological Agency: Formulation and preliminary results. *Quart. J. Roy. Meteor. Soc.*, **131**, 3465–3475, <https://doi.org/10.1256/qj.05.132>.
- Ingleby, N. B., A. C. Lorenc, K. Ngan, F. Rawlins, and D. Jackson, 2013: Improved variational analyses using a nonlinear humidity control variable. *Quart. J. Roy. Meteor. Soc.*, **139**, 1875–1887, <https://doi.org/10.1002/qj.2073>.
- Jones, C. D., and B. Macpherson, 1997: A latent heat nudging scheme for the assimilation of precipitation data into an operational mesoscale model. *Meteor. Appl.*, **4**, 269–277, <https://doi.org/10.1017/S1350482797000522>.
- Rogers, R., and M. Yau, 1989: *A Short Course in Cloud Physics*. Butterworth-Heinemann, 304 pp.
- Saito, K., 2012: The JMA nonhydrostatic model and its applications to operation and research. *Atmospheric Model Applications*, I. Yucel, Ed., InTech, 85–110.
- , and Coauthors, 2006: The operational JMA nonhydrostatic mesoscale model. *Mon. Wea. Rev.*, **134**, 1266–1298, <https://doi.org/10.1175/MWR3120.1>.
- , J. Ishida, K. Aranami, T. Hara, T. Segawa, M. Narita, and Y. Honda, 2007: Nonhydrostatic atmospheric models and operational development at JMA. *J. Meteor. Soc. Japan*, **85B**, 271–304, <https://doi.org/10.2151/jmsj.85B.271>.
- Stillier, O., and S. P. Ballard, 2009: Efficient moist physics schemes for data assimilation. I: Large-scale clouds and condensation. *Quart. J. Roy. Meteor. Soc.*, **135**, 707–720, <https://doi.org/10.1002/qj.400>.