Quarterly Journal of the Royal Meteorological Society Q. J. R. Meteorol. Soc. 143: 785–797, January 2017 B DOI:10.1002/qj.2964

RMetS
Royal Meteorological Society

The effect of improved ensemble covariances on hybrid variational data assimilation

N. E. Bowler,* A. M. Clayton, M. Jardak, P. M. Jermey, A. C. Lorenc, M. A. Wlasak, D. M. Barker, G. W. Inverarity and R. Swinbank

Met Office, Exeter, UK

*Correspondence to: N. E. Bowler, Met Office, Fitzroy Road, Exeter EX1 3PB, UK. E-mail: neill.bowler@metoffice.gov.uk

Hybrid four-dimensional ensemble-variational (4DEnVar) data assimilation is a method which avoids using a linear and adjoint model by relying on an input ensemble to propagate analysis increments in time. Previous studies have shown that hybrid 4DEnVar performs worse than hybrid four-dimensional variational (4D-Var) assimilation. Given hybrid 4DEnVar's heavy reliance on the ensemble, this comparison may be affected by the quality of the input ensemble. Here we investigate how improvements to the ensemble system affect hybrid 4D-Var and how they affect the comparison with hybrid 4DEnVar.

Using the Met Office's operational ensemble generation scheme (the ensemble transform Kalman filter, ETKF) it is found that hybrid 4D-Var gains little benefit from using an enlarged ensemble as input (176 as opposed to 23 members). By contrast, hybrid 4DEnVar benefits more from the increased ensemble size, and it benefits further when the weighting given to the ensemble covariance is increased. Both data assimilation methods benefit when the input ensemble is changed from using the ETKF to using an ensemble of 4DEnVars. Both schemes also show further benefit when a large ensemble (200 members) of 4DEnVars is used, and when a large weight is given to the covariance information from this ensemble. Thus, improving the ensemble covariance used in assimilation (ensemble generation method and ensemble size) and increasing its weight can have substantial benefits.

Given that both hybrid 4D-Var and hybrid 4DEnVar benefit from improvements to the input ensemble, the relative performance is largely unaffected by the ensemble changes and hybrid 4D-Var performs better than hybrid 4DEnVar for all input ensembles.

Key Words: ensemble; data assimilation; numerical weather prediction; ensemble Kalman filter

Received 8 February 2016; Revised 21 October 2016; Accepted 28 October 2016; Published online in Wiley Online Library 8 February 2017

1. Introduction

Variational data assimilation (DA) methods have been widely used for large-scale numerical weather prediction (NWP) because they make effective use of a wide variety of observations. Information from a prior forecast (known as the background) is combined with observations, using a weighting determined by the errors in each. Four-dimensional variational assimilation (4D-Var), a method which incorporates a linear model and its adjoint into the variational algorithm (Talagrand, 2010), has been successfully implemented in several centres (Rabier, 2005) to correctly allow for the actual time of observations. This last effect was studied by Lorenc and Rawlins (2005), and a large benefit was measured in experiments with a modern system by Lorenc *et al.* (2015). A disadvantage of 4D-Var is the cost of maintaining and running the linear and adjoint models – these costs are expected to increase on future massively parallel computers.

Recently, there has been much interest in improving the static background error covariance used in traditional variational methods by incorporating covariances from a current ensemble of forecasts. This is the basis of hybrid 4D-Var which was implemented operationally by Clayton *et al.* (2013); their introduction gives earlier references. However hybrid 4D-Var retains the costs of the linear and adjoint models.

Four-dimensional ensemble-variational data assimilation (4DEnVar) was introduced by Liu *et al.* (2008) to avoid these costs by using the ensemble at several times. It was tested first in an NWP model by Buehner *et al.* (2010a,b). We have previously made a clean comparison of hybrid 4DEnVar with the operational hybrid 4D-Var in the Met Office system (Lorenc *et al.*, 2015), finding that hybrid 4D-Var performed better. We concluded that one of the main reasons is that we currently give little weight to the flow-dependent covariances coming from our ensemble prediction system (EPS). If we improve the generation of the ensemble perturbations, then it is hoped that we will be able to give a higher weight to this information and the performance of the new hybrid 4DEnVar assimilation method will be improved relative to the operational

system. This was the prime motivation of the work reported here.

In the recent past, Environment and Climate Change Canada have moved from using 4D-Var, which did not use covariance information from the ensemble, to hybrid 4DEnVar as their operational DA scheme (Buehner *et al.*, 2015). Clayton *et al.* (2013) demonstrated considerable benefit from the use of ensemble covariance information in the Met Office's operational hybrid 4D-Var. To be able to compare with the Canadian experience, and to update the results of Clayton *et al.* (2013) to an improved ensemble, we will also investigate the performance of non-hybrid 4D-Var relative to the other methods.

In a companion article (Bowler *et al.*, 2017) we investigate the use of an ensemble of 4DEnVars (En-4DEnVar) to replace the Met Office's Ensemble Transform Kalman Filter (ETKF) system (Bishop *et al.*, 2001; Bowler *et al.*, 2008), which has been operational since September 2008. The main focus of this study is to examine the impact of changing the ensemble information used in the DA. We look at both changing the ensemble size and the system used to generate the ensemble perturbations.

Wang *et al.* (2013) compared a (non-hybrid) 3D-Var scheme with a fully-ensemble 3DEnVar scheme and a hybrid 3DEnVar. They found that the fully-ensemble 3DEnVar performed better than 3D-Var and was not improved by the use of hybrid background-error covariances. They concluded that a hybrid was not needed because they ran the ensemble at the same resolution as the DA and used a large ensemble size (80 members). Kleist and Ide (2015a,b) used a similar system within an observing system simulation experiment to examine this issue, but with an ensemble run at a lower resolution than the DA. They found that hybrid 3DEnVar performed much better than (non-hybrid) 3D-Var and very similarly to fully-ensemble 3DEnVar. Conversely, they found that hybrid 4DEnVar performed better than fully-ensemble 4DEnVar. They suggest that this difference may depend on the way in which imbalance is dealt with in their system.

Poterjoy and Zhang (2015, 2016) compared the hybrid four-dimensional variational methods which we are studying. Their second article found that hybrid 4D-Var performed better in a modern hurricane forecasting system. They hypothesized that this was because hybrid 4DEnVar needs to localize covariances in time.

1.1. 4D data assimilation methods

The DA methods used in this article are all based on four-dimensional variational minimizations. They are described in more detail in Clayton *et al.* (2013) and Lorenc *et al.* (2015), so will only be described briefly here. The process of DA seeks the optimal analysis state which combines information from a background forecast (which embodies all the information from previous cycles of the assimilation system and physical constraints) with the latest observational data. Variational methods achieve this by minimising the cost function

$$J\left(\underline{\delta\mathbf{x}}\right) = \frac{1}{2}\underline{\delta\mathbf{x}}^{\mathrm{T}}\underline{\mathbf{P}}^{-1}\underline{\delta\mathbf{x}} + \frac{1}{2}\left(\underline{\mathbf{y}} - \underline{\mathbf{y}}^{\mathrm{o}}\right)^{\mathrm{T}}\underline{\mathbf{R}}^{-1}\left(\underline{\mathbf{y}} - \underline{\mathbf{y}}^{\mathrm{o}}\right), \tag{1}$$

where $\underline{\delta\mathbf{x}}$ is the 4D increment applied to the background state over the assimilation window and $\mathbf{P}$ is the error covariance matrix of this background state. Terms written with underlines indicate that the the associated quantities are distributed over a window of time. $\underline{\mathbf{y}}^{\mathrm{o}}$ are the observations spanning the assimilation window and $\underline{\mathbf{R}}$ is the block-diagonal observation-error covariance matrix, with the blocks representing the covariances at successive observation times. The model equivalents of the observations over the time window ($\underline{\mathbf{y}}$) are calculated as accurately as possible (Lorenc *et al.*, 2000) using

$$\underline{\mathbf{y}} = \underline{H}\{\underline{M}(\mathbf{x}^{\mathrm{b}}) + \underline{\delta\mathbf{x}}\}. \tag{2}$$

Note that the observation operator $\underline{H}$ is potentially nonlinear and can be applied to the incremented state, not just the background state. Finally, the superscript T denotes matrix transposition or, in the case of vectors, transformation from a column vector to a row vector.

For large NWP systems it is not feasible to represent $\underline{\mathbf{P}}$ directly, so variational methods define $\underline{\delta\mathbf{x}}$ and $\underline{\mathbf{P}}$ such that $\underline{\delta\mathbf{x}}^{\mathrm{T}}\underline{\mathbf{P}}^{-1}\underline{\delta\mathbf{x}}$ simplifies. The methods described below differ in the way they do this.

1.2. Hybrid 4D-Var

'Strong-constraint' 4D-Var uses an approximate tangent linear model and its adjoint to propagate the analysis increments in time (Rawlins *et al.*, 2007), and ignores its errors. This linear model is denoted $\underline{\mathbf{M}}$ and is related to the nonlinear model using

$$\underline{M}(\mathbf{x}^{\mathrm{b}} + \delta\mathbf{x}) \simeq \underline{M}(\mathbf{x}^{\mathrm{b}}) + \underline{\mathbf{M}}\delta\mathbf{x}. \tag{3}$$

Thus $\underline{\delta\mathbf{x}} = \underline{\mathbf{M}}\delta\mathbf{x}$ is the perturbation trajectory through the assimilation window, starting from the increment $\delta\mathbf{x}$ at the beginning of the window. Thanks to the strong-constraint assumption, the errors through the window can be related to those at the beginning by $\underline{\mathbf{P}} = \underline{\mathbf{M}}\mathbf{P}\underline{\mathbf{M}}^{\mathrm{T}}$.

Traditional 4D-Var algorithms represent a time-invariant estimate of the background-error covariance matrix (the so-called climatological covariance) using a static set of transform operators

$$\mathbf{B}_{\mathrm{c}} = \mathbf{U}\mathbf{U}^{\mathrm{T}}, \tag{4}$$

where the $\mathbf{U}$ operator transforms a set of control variables in the vector $\mathbf{v}$, each drawn from a standard normal distribution, into the increment through $\delta\mathbf{x} = \mathbf{U}\mathbf{v}$. On the other hand, ensemble methods use a set of localized perturbations to represent the background-error covariance

$$\mathbf{B}_{\mathrm{e}} = \mathbf{C} \circ \mathbf{X}\mathbf{X}^{\mathrm{T}}, \tag{5}$$

where $\mathbf{C}$ is the localization matrix and $\mathbf{X} = \left[\mathbf{x}'_1, \mathbf{x}'_2, \ldots, \mathbf{x}'_N\right]/\sqrt{N-1}$ gives the normalized ensemble perturbations, and $\mathbf{x}'_i = \mathbf{x}_i - \overline{\mathbf{x}}$ is the perturbation of the ith ensemble member $\mathbf{x}_i$ from the ensemble mean $\overline{\mathbf{x}}$; we refer to these perturbations as error modes. The error modes used in this article come from ensemble systems described by Bowler *et al.* (2017). The element-wise Schur product is denoted by $\circ$ and N is the ensemble size. Neither of these representations of the background errors is perfect so it is common practice to use a weighted combination of both (Hamill and Snyder, 2000)

$$\mathbf{P} = \beta_{\mathrm{c}}^2\mathbf{U}\mathbf{U}^{\mathrm{T}} + \beta_{\mathrm{e}}^2\mathbf{C} \circ \mathbf{X}\mathbf{X}^{\mathrm{T}}, \tag{6}$$

where β_{c}^2 and β_{e}^2 are the weights. If we assume that both components give independent estimates of the true background-error covariance, then we would expect

$$\beta_{\mathrm{c}}^2 + \beta_{\mathrm{e}}^2 = 1. \tag{7}$$

However, Clayton *et al.* (2013) found that having $\beta_{\mathrm{c}}^2 = 0.8$ and $\beta_{\mathrm{e}}^2 = 0.5$ gave best performance. The reason is that the climatological covariances as implemented in the current Met Office system are underspread* for potential temperature and pressure, and over-weighting the static component is a simple way to account for this deficiency. We could regard this as setting $\beta_{\mathrm{c}}^2 = 0.62$ and $\beta_{\mathrm{e}}^2 = 0.38$, with an inflation of 1.3 to account

*Underspread is normally defined as an ensemble spread which is smaller than the root-mean-square error of the ensemble mean. In the context of climatological covariances underspread means that the estimated variances are less than the typical errors in the background state.

for the underspread. Following further tuning, the Met Office's operational system has used $\beta_c^2 = 1$ and $\beta_e^2 = 0.3$ since January 2013, and this is used as the starting point for our experiments.

When using the ensemble information, we choose to perform the localization in variables which have been transformed to have errors which are less correlated than the prognostic variables in our model. This may also reduce the level of imbalance in the analysis (Kepert, 2009). Using this transformation is in contrast to other studies (e.g. Buehner *et al.*, 2015). Let $\mathbf{T}_p$ denote the operator of such a transformation with right inverse $\mathbf{U}_p$, i.e. $\mathbf{T}_p\mathbf{U}_p = \mathbf{I}$, the identity matrix. When applying $\mathbf{T}_p$, the modified hybrid background-error covariance matrix becomes

$$\mathbf{P} = \beta_c^2\mathbf{U}\mathbf{U}^{\mathrm{T}} + \beta_e^2\mathbf{U}_p\left(\mathbf{C} \circ \mathbf{T}_p\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{T}_p^{\mathrm{T}}\right)\mathbf{U}_p^{\mathrm{T}}. \tag{8}$$

Rather than finding $\delta\mathbf{x}$ directly, it is more computationally efficient to solve for control variables. For hybrid methods we combine two sources of background-error information and so we have two control vectors: $\mathbf{v}$ is related to the static component and $\mathbf{w}$ to the ensemble component. Thus the analysis increment at the start of the window is

$$\delta\mathbf{x} = \beta_c\mathbf{U}\mathbf{v} + \beta_e\mathbf{U}_p\sum_i \mathbf{L}\mathbf{w}_i \circ \mathbf{T}_p\mathbf{x}_i', \tag{9}$$

where $\mathbf{w}_i$ is the control vector for ensemble member i and $\mathbf{L}$ is a Cholesky factor of the localization matrix $\mathbf{C} = \mathbf{L}\mathbf{L}^{\mathrm{T}}$. The principal difference between 4D-Var and 4DEnVar is how they propagate this increment to the observation time.

To construct the hybrid 4D-Var cost function, we combine Eqs (1)–(3), (8) and (9) to get

$$J_{4\mathrm{DVar}}(\mathbf{v},\mathbf{w}) = \frac{1}{2}\mathbf{v}^{\mathrm{T}}\mathbf{v} + \frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} + J_c + \frac{1}{2}(\underline{\mathbf{y}}^{\mathrm{o}} - \underline{\mathbf{y}})^{\mathrm{T}}\underline{\mathbf{R}}^{-1}(\underline{\mathbf{y}}^{\mathrm{o}} - \underline{\mathbf{y}}). \tag{10}$$

The J_c term, similar to that of Gauthier and Thépaut (2001), penalizes high-frequency noise in the model trajectory to improve balance in the model forecast. Recent tests have indicated that the use of a J_c term has only small impacts on hybrid 4D-Var. The use of a hybrid background-error covariance has been effective in 4D-Var, leading to a clear improvements in forecasts over non-hybrid 4D-Var (Clayton *et al.*, 2013).

1.3. Hybrid 4DEnVar

Since each ensemble forecast is run throughout the following DA window, we can use these forecasts instead of a linear model. We define the trajectories for the model perturbations as $\underline{\mathbf{X}} = \left[\underline{\mathbf{x}}_1', \underline{\mathbf{x}}_2', \ldots, \underline{\mathbf{x}}_N'\right]/\sqrt{N-1}$. In this case the four-dimensional analysis increment may be written as

$$\underline{\delta\mathbf{x}} = \beta_c\underline{\mathbf{I}}\mathbf{U}\mathbf{v} + \beta_e\underline{\mathbf{U}}_p\sum_i(\underline{\mathbf{I}}\mathbf{L}\mathbf{w}_i) \circ \underline{\mathbf{T}}_p\underline{\mathbf{x}}_i', \tag{11}$$

where underlines have been added to the $\mathbf{T}_p$ and $\mathbf{U}_p$ operators because of their time-dependence by virtue of being functions of the ensemble mean. This is the most common and simplest form of 4DEnVar, with the spatial localization matrix (like that used in hybrid 4D-Var) applied to a four-dimensional ensemble covariance; to achieve this the identity operator $\underline{\mathbf{I}}$ copies its argument to all times. Despite this constant localization, the second term in this expression differs at each time in the assimilation window because the ensemble is propagated by the nonlinear model, and parameter transform operators are time-dependent. The first term, on the other hand, is simply a copy of the climatological increment to the different times in the window. Thus, the increment from this portion does not propagate with the dynamics of the model, unlike in hybrid 4D-Var, where the value at the start of the time window is evolved by the linear model.

The hybrid 4DEnVar cost function is very similar to the hybrid 4D-Var cost function, except that the analysis increment is now four-dimensional:

$$J_{4\mathrm{DEnVar}}(\mathbf{v},\mathbf{w}) = \frac{1}{2}\mathbf{v}^{\mathrm{T}}\mathbf{v} + \frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} + \frac{1}{2}(\underline{\mathbf{y}}^{\mathrm{o}} - \underline{\mathbf{y}})^{\mathrm{T}}\underline{\mathbf{R}}^{-1}(\underline{\mathbf{y}}^{\mathrm{o}} - \underline{\mathbf{y}}). \tag{12}$$

The 4D-Var analysis increment may also be written in a four-dimensional form using the linear model, although it is rarely expressed in this way. The J_c term as coded in hybrid 4D-Var uses the linear perturbation forecast model (Rawlins *et al.*, 2007) and as such cannot be used in 4DEnVar. The main effect of J_c is to control the temporal high-frequency noise in the 4D-Var experiments. To achieve a similar effect in our hybrid 4DEnVar experiments, we use the 4DIAU (Lorenc *et al.*, 2015) as a replacement for J_c. 4DIAU is the best initialization we have for hybrid 4DEnVar, but is not very effective for hybrid 4D-Var.

2. Experiment details

2.1. Configurations

The experiments are split into two sets, based on substantially different NWP configurations and different seasons.

The first set was run a year and a half before the second, and was based on the global NWP configuration made operational at the Korea Meteorological Administration (KMA) in June 2013. The Met Office Unified Model (MetUM) forecasts were run on an N512L70 grid[†] with a Global Atmosphere, GA3.1 (Walters *et al.*, 2011) model configuration (New Dynamics; Davies *et al.*, 2005). DA was performed by (hybrid or non-hybrid) 4D-Var or hybrid 4DEnVar, using data from an ETKF-based ensemble (Bowler *et al.*, 2008) using an N320L70 grid, again with a GA3.1 model configuration. The 4D-Var analyses were computed using an N216L70 grid and the hybrid 4DEnVar analyses used an N320L70 grid. Two ensemble experiments were run for the period 0000 UTC 1 August 2012 to 1200 UTC 31 August 2012, one with 23 and one with 176 (perturbed) members, centring around global analyses from KMA's operational system. These experiments, described in the companion article (Bowler *et al.*, 2017), saved the data required for the subsequent hybrid 4D-Var and hybrid 4DEnVar experiments. The DA experiments were started at 0000 UTC on 8 August 2012 to allow a week's spin-up for the ensemble, and were also run to 1200 UTC on 31 August 2012 – 23 days in total. All forecasts were verified, so there was no allowance made for spin-up in the DA experiments. Moreover, forecasts that extended beyond the end of the experiment period were not verified, so that, for example, only 19 days' worth of T + 120 forecasts (denoting 120 h after the nominal cycle time) were verified for each experiment, albeit from both 0000 UTC and 1200 UTC analyses. Thus, the verification scores are affected by a large amount of statistical noise, and by spin-up. For this reason – plus the somewhat outdated configuration – most of our analysis will focus on the second set of experiments, which span a much longer time period and include a much wider range of configurations. Unless stated otherwise, results will be from the second set of experiments and results from the first set will only be discussed where they add useful supporting information. In plots, the first set of experiments is denoted by 'Aug12_ND'.

[†]N512 refers to a regular latitude–longitude grid with double the specified number points around each latitude circle and 1.5 times the specified number plus one points on each line of longitude from pole to pole for the New Dynamics dynamical core. Thus an N512 grid has 1024 × 768 grid points. L70 refers to the number of model levels in the vertical. All DA is done using the New Dynamics grid. The ENDGAME dynamical core uses the same notation but has a different arrangement of grid points with one fewer grid point on each line of longitude from pole to pole.

The second set of experiments was run at both KMA and ECMWF (European Centre for Medium-Range Weather Forecasts) and was based on a lower-resolution configuration, but with an updated forecast model using the ENDGAME dynamical core (Wood *et al.*, 2014). The same grid notation is used as for the New Dynamics dynamical core, but it should be noted that the grid-staggering is different, meaning there is one fewer row for temperature and pressure variables. The set-up was chosen to mimic the configuration used in the Met Office's parallel suite 35, which became operational on 3 February 2015. The deterministic forecasts were run on an N320L70 grid and, as for the first set of experiments, 4D-Var used an N216L70 grid. We ran three ensemble experiments: an ETKF-based ensemble with 44 members, and 4DEnVar-based ensembles with 44 and 200 members. These are listed as experiments 18, 20 and 17, respectively in Bowler *et al.* (2017). These ensembles were all run using N216L70 grids, and hence so were the hybrid 4DEnVar analyses. Thus, unlike for the first set of experiments, there was no resolution advantage for hybrid 4DEnVar. Both the ensembles and the DA experiments were run for the period 0000 UTC 24 January 2014 to 0000 UTC 14 March 2014. To allow for spin-up, only forecasts valid from 0000 UTC 31 January 2014 were verified, including those which extended beyond the end of the experiment period, giving 42 days' worth of forecasts for each forecast range, from both 0000 UTC and 1200 UTC analyses. In plots, this set of experiments is denoted by 'Feb14_EG'.

Both sets of experiments consider different weightings to the ensemble and static covariances. The operational hybrid 4D-Var system uses a weighting of $\beta_c^2 = 1$ to the climatological (static) background-error covariances, and a weighting of $\beta_e^2 = 0.3$ to the ensemble-based background-error covariances. The experiments primarily focus on the effect of increasing the weight to the ensemble covariances. The static covariances are derived from ECMWF's EDA (ensemble of DAs), which runs an ensemble of 4D-Vars (Fisher, 2003). These are then evolved using the MetUM before being used to calibrate the static covariances. The dependence on the ECMWF model, however, means that the resulting covariances have different characteristics to the MetUM forecast model used in these experiments. Future work will consider deriving the static covariance statistics using the Met Office EPS.

The update of sea-surface temperature and sea-ice is performed operationally using the Operational Sea Surface Temperature and Sea Ice Analysis (OSTIA) system (Donlon *et al.*, 2012). The analyses from this system are archived, and then used directly in the analysis, and this system was adopted for both sets of experiments. The land-surface fields (soil moisture and temperature) were updated using a nudging approach (Dharssi *et al.*, 2011) for the first set of experiments, and using an extended Kalman filter (EKF; Candy, 2014) for the second set.

As noted above, the 4D-Var analyses were run on the N216L70 grid for both experiment sets. These included an initial minimization using an N108L70 grid with 30 iterations to provide a starting point and preconditioning for the higher-resolution minimization. This double-resolution approach improves computational efficiency but has no overall effect on the accuracy of the analysis and is like our current operational 4D-Var system. As such it is a good baseline for testing the accuracy of 4DEnVar analyses, which do not use this technique. For the second set of experiments, the N216L70 part of the analysis was run with a 12 min timestep in the linear perturbation forecast model (Rawlins *et al.*, 2007) (rather than 20 min) and the minimization used 40 rather than 30 iterations.

The hybrid 4DEnVar analyses use only a single minimization on the N216L70 grid using the same configuration with 100 minimization iterations in the first set of experiments and 60 in the second. Despite the improved efficiency of the double-resolution minimization strategy, execution times for hybrid 4D-Var analyses are much longer than for hybrid 4DEnVar because of the high cost of the linear and adjoint model integrations. To make a fair cost comparison between hybrid 4D-Var and hybrid 4DEnVar, we switched hybrid 4D-Var to the single-resolution 60-iteration minimization used by hybrid 4DEnVar in the second set of experiments, and ran some test analyses using equalized computer resources. The total execution time for hybrid 4DEnVar was 17% that of hybrid 4D-Var with 44 members, and 30% with 200 members. However, 4DEnVar spends a much larger proportion of its execution time reading and pre-processing the ensemble data, so if we just consider the time spent in the main minimization loop, the percentages drop to 11 and 18% respectively. At higher resolutions the timing advantage for hybrid 4DEnVar increases due to the relatively poor scaling of the 4D-Var linear and adjoint models with resolution. However, the memory requirements of hybrid 4DEnVar are much higher, which may restrict practical ensemble sizes and resolutions on some computer systems.

The ensemble localization settings were also varied between experiments, as will be explained below in context. Because the 4DEnVar code does not currently support dust assimilation, this was turned off in the 4D-Var experiments. All the experiments were run with one-way coupling – the ensemble was run first and provided information to the DA, but it received no information from the DA. Others (e.g. Wang *et al.*, 2013) have experimented with two-way coupling, but this is not tested here.

The code used in these experiments was based on standard versions hosted in the Met Office repositories. Since the first set of experiments was run much earlier, it was based on older code versions: SURF 18.5 for surface DA, OPS 27.2 for observation processing, UM 7.9 for the forecast, VAR 27.2 for atmospheric DA and VER 14.1 for verification. The second set was based on SURF 31.2, OPS 31.1, UM 8.5, VAR 32.0 and VER 17.1. Latest versions of the code and documentation are stored at http://code.metoffice.gov.uk/ (accessed 7 December 2016).

Within each experiment set, the configurations differ along four 'dimensions':
(i) assimilation algorithm (4D-Var or 4DEnVar);
(ii) input ensemble data;
(iii) covariance weightings; and
(iv) localization type.
Thus for plot labelling we have given each experiment an ID consisting of four dot-separated fields, which identify the variants used along each dimension. Tables 1 and 2 give the configurations that were run for the first and second sets of experiments, respectively.

2.2. *Verification approach*

To assess the quality of the forecasts, we rely mainly on the root-mean-square (RMS) error of the forecast. This is calculated over various regions of the globe and has been assessed against observations and ECMWF analyses (interpolated to a 1.5° latitude–longitude grid). When verifying against observations, surface observations are used to measure pressure at mean sea-level (PMSL), sondes are used to measure geopotential height, aircraft observations are used for winds at 250 hPa and satellite-derived atmospheric motion vectors (AMVs; Schmetz *et al.*, 1993) for winds at 850 hPa. The first set of experiments did not calculate verification against aircraft or AMV observations, so winds are verified against sondes for these experiments.

The graphics shown here mainly report the change in RMS error (RMSE) between two experiments. These scores are displayed graphically as an array of triangles. An upward-pointing green (downward-pointing blue) triangle indicates that the experiment being tested has a lower (higher) RMSE than the control experiment against which it is being compared. The area of each triangle is proportional to the percentage change in the RMSE. The maximum symbol size in the plot represents a change in RMSE of 5% or more. We will often then compose the changes in RMSE into an NWP index (Appendix) which provides a single number to summarize the quality of the experiment. Clearly,

Table 1. Summary of experiments run using the New Dynamics MetUM.

Experiment ID	Var type	Ens. type	Ens. size	β_c^2 (%)	β_e^2 (%)	Localization type	Run at
4DVar.non-hybrid	4D-Var (non-hybrid)	–	–	100	0	–	KMA
4DVar.etkf23.100c30e.OpLoc	4D-Var	ETKF	23	100	30	OpLoc	KMA
4DVar.etkf23.80c50e.OpLoc	4D-Var	ETKF	23	80	50	OpLoc	KMA
4DVar.etkf23.30c70e.OpLoc	4D-Var	ETKF	23	30	70	OpLoc	KMA
4DVar.etkf176.100c30e.OpLoc	4D-Var	ETKF	176	100	30	OpLoc	KMA
4DVar.etkf176.80c50e.OpLoc	4D-Var	ETKF	176	80	50	OpLoc	KMA
4DVar.etkf176.30c70e.OpLoc	4D-Var	ETKF	176	30	70	OpLoc	KMA
4DEnVar.etkf23.80c50e.OpLoc	4DEnVar	ETKF	23	80	50	OpLoc	KMA
4DEnVar.etkf23.30c70e.OpLoc	4DEnVar	ETKF	23	30	70	OpLoc	KMA
4DEnVar.etkf176.80c50e.OpLoc	4DEnVar	ETKF	176	80	50	OpLoc	KMA
4DEnVar.etkf176.30c70e.OpLoc	4DEnVar	ETKF	176	30	70	OpLoc	KMA

The experiment ID describes the variational formulation, input ensemble type and size, weights to climatological and ensemble covariance components and the localization type, e.g. 4DVar.etkf23.30c70e.OpLoc refers to a hybrid 4D-Var experiment driven by a 23 member ETKF ensemble, applying 30% weighting to the climatological covariance and 70% to the ensemble covariance and using operational-like localization. OpLoc localization relaxes to the climatological covariance above 21 km.

Table 2. Summary of experiments run using the ENDGAME MetUM.

Experiment ID	Var type	Ens. type	Ens. size	β_c^2 (%)	β_e^2 (%)	Localization type	Run at
4DVar.non-hybrid	4D-Var (non-hybrid)	–	–	100	0	–	KMA
4DVar.etkf44.100c30e.NewLoc	4D-Var	ETKF	44	100	30	NewLoc	ECMWF
4DVar.etkf44.100c30e.OpLoc	4D-Var	ETKF	44	100	30	OpLoc	ECMWF
4DVar.etkf44.30c70e.NewLoc	4D-Var	ETKF	44	30	70	NewLoc	ECMWF
4DVar.envar44.100c30e.NewLoc	4D-Var	En-4DEnVar	44	100	30	NewLoc	ECMWF
4DVar.envar44.30c70e.NewLoc	4D-Var	En-4DEnVar	44	30	70	NewLoc	ECMWF
4DVar.envar44.30c140e.NewLoc	4D-Var	En-4DEnVar	44	30	140	NewLoc	ECMWF
4DVar.envar44.0c100e.BigLoc	4D-Var	En-4DEnVar	44	0	100	BigLoc	ECMWF
4DVar.envar44.0c100e.NewLoc	4D-Var	En-4DEnVar	44	0	100	NewLoc	ECMWF
4DVar.envar200.100c30e.NewLoc	4D-Var	En-4DEnVar	200	100	30	NewLoc	KMA
4DVar.envar200.100c30e.OpLoc	4D-Var	En-4DEnVar	200	100	30	OpLoc	KMA
4DVar.envar200.30c70e.NewLoc	4D-Var	En-4DEnVar	200	30	70	NewLoc	KMA
4DVar.envar200.30c70e.OpLoc	4D-Var	En-4DEnVar	200	30	70	OpLoc	KMA
4DEnVar.envar44.0c100e.BigLoc	4DEnVar	En-4DEnVar	44	0	100	BigLoc	ECMWF
4DEnVar.envar44.0c100e.NewLoc	4DEnVar	En-4DEnVar	44	0	100	NewLoc	ECMWF
4DEnVar.envar44.30c70e.NewLoc	4DEnVar	En-4DEnVar	44	30	70	NewLoc	ECMWF
4DEnVar.envar200.100c30e.NewLoc	4DEnVar	En-4DEnVar	200	100	30	NewLoc	KMA
4DEnVar.envar200.100c30e.OpLoc	4DEnVar	En-4DEnVar	200	100	30	OpLoc	KMA
4DEnVar.envar200.30c70e.NewLoc	4DEnVar	En-4DEnVar	200	30	70	NewLoc	KMA
4DEnVar.envar200.30c70e.OpLoc	4DEnVar	En-4DEnVar	200	30	70	OpLoc	KMA

The experiment IDs and OpLoc are defined as in Table 1. NewLoc refers to new localization, using Gaspari–Cohn functions for vertical localization, and no relaxation to the climatological covariance at upper levels. BigLoc is the same as NewLoc, but with a doubled localization length-scale in the horizontal, and a one-third increase in the vertical.

using a single number to summarize the entire performance of an experiment carries risks and should be interpreted with caution.

Unless otherwise stated, all the verification plots use verification against ECMWF analyses for forecasts which are initialized between 0000 UTC on 31 January and 0000 UTC on 14 March 2014. Thus, we consider forecasts which are valid up to 0000 UTC on 19 March 2014.

3. Hybrid versus non-hybrid 4D-Var with operational-like settings

To provide a standard against which to compare other changes below, Figure 1 shows the impact of moving from non-hybrid 4D-Var to an operational-like hybrid 4D-Var configuration; i.e. a configuration that uses an ETKF-based ensemble with a weight of $\beta_c^2 = 1.0$ given to the climatological covariance and a weight of $\beta_e^2 = 0.3$ to the ensemble covariance, using the localization scheme that has been operational since July 2011. We see a very clear and general improvement with the RMSE being lower for the hybrid configuration across most variables. The average change is approximately a 1.3% reduction in the RMSE. This compares with a 3.6% reduction in RMSE when comparing hybrid 3D-Var (first guess at appropriate time; Lorenc and Rawlins, 2005) with hybrid 4D-Var (Lorenc *et al.*, 2015). The same general improvement (not shown) is seen for the older experiment configuration but with a smaller overall impact, presumably due to the ensemble having 23 rather than 44 members.

4. Effects of changing the ensemble and its usage in hybrid 4D-Var

In this section we look at the effects of changing the ensemble system – the basic technique (ETKF or En-4DEnVar) and the ensemble size – and changing the way the ensemble data are used in hybrid 4D-Var, by modifying the localization scheme and the covariance weightings. Results from the 4DEnVar experiments will be discussed later, in section 5.

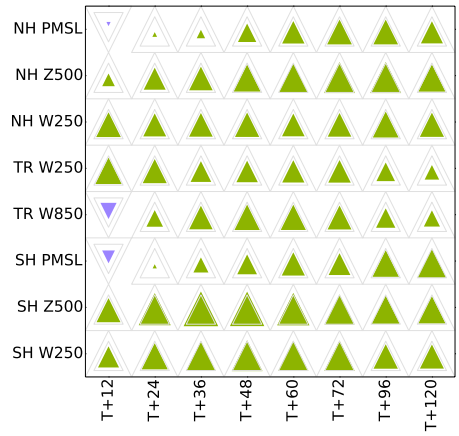

Figure 1. Relative change in RMSE for NWP index fields after moving to an operational-like hybrid configuration from non-hybrid 4D-Var (i.e. comparing 4DVar.etkf44.100c30e.OpLoc with 4DVar.non-hybrid). Verification is against ECMWF analyses from 0000 UTC on 31 January 2014 to 0000 UTC on 19 March 2014. The legend on the left refers to forecast components of mean sea-level pressure (PMSL), wind (W) and geopotential height (Z) at 250, 500 and 850 hPa pressure levels in the Northern Hemisphere (NH), Tropics (TR) and Southern Hemisphere (SH). The hollow grey triangles show the size for a 2 and 5% change in RMSE. The biggest difference that can be plotted is 5%. The upward-pointing green (downward-pointing blue) triangles are plotted if the change reduced (increased) the RMSE. [Colour figure can be viewed at wileyonlinelibrary.com].

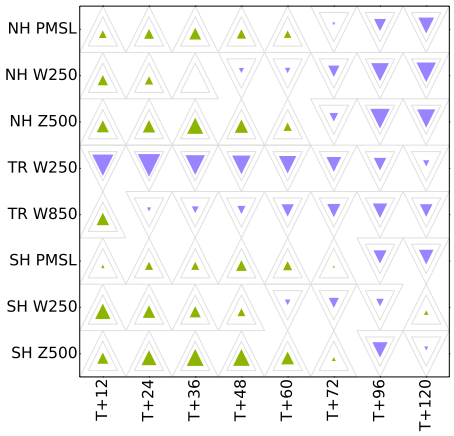

Figure 2. As Figure 1, but showing change from reducing the horizontal localization length-scale and using hybrid covariances throughout the depth of the model (i.e. comparing 4DVar.etkf44.100c30e.NewLoc with 4DVar.etkf44.100c30e.OpLoc). [Colour figure can be viewed at wileyonlinelibrary.com].

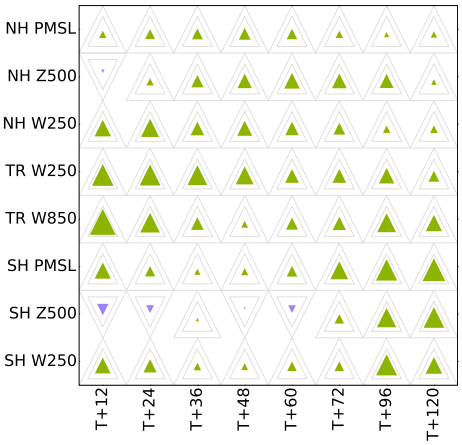

Figure 3. As Figure 1, but showing change from using En-4DEnVar rather than ETKF error modes when a high weight is given to the climatological covariance (i.e. comparing 4DVar.envar44.100c30e.NewLoc with 4DVar.etkf44.100c30e.NewLoc). [Colour figure can be viewed at wileyonlinelibrary.com].

4.1. *Localization changes*

In previous ensemble-DA experiments, we experienced problems with using ensemble information near the top of the model. If a large analysis increment near the top of the model was used, then subsequent model runs occasionally failed. Thus our operational configuration uses ensemble information only below 21 km, with the weight given to the ensemble covariance reducing smoothly between 16 and 21 km (Clayton *et al.*, 2013). However, this might cause problems in experiments where a high weight is given to the ensemble covariance in the lower part of the model, since there would then be a sharp transition between using mostly ensemble covariances to wholly climatological covariances. Therefore, in most of our experiments we have kept the same weight to the ensemble covariance throughout the depth of the model.

Additionally, the operational system uses broad horizontal and vertical localization length-scales which we found to be detrimental in the ensemble experiments (Bowler *et al.*, 2017). Therefore in most of these experiments, we used the same tighter localization as in the ensemble experiments. Specifically, for the experiments using 44 ensemble members the Gaussian horizontal length-scale was set to 600 km (reduced from 1200 km) and the vertical localization is based on the function of Gaspari and Cohn (1999) with a cut-off length-scale of 1.5 scale-heights (changed from a scheme based on vertical correlations of streamfunction). One scale-height (or e-folding distance) is the distance over which the pressure changes by a factor of $e \approx 2.71828$.

Figure 2 shows the impact of these localization changes, in the context of experiments using an ETKF-based ensemble with 44 members, giving a high weight to the climatological covariance. This tests using the modified horizontal and vertical localization, and using the ensemble covariance throughout the depth of the model. We see that the changes have broadly positive impacts in the Extratropics at short lead times, and negative impacts in the Tropics and at longer lead times. Extended verification also indicates that the performance is worse higher up in the model, and generally better at lower levels. Length-scales tend to be longer in the stratosphere than in the troposphere, so this may indicate a disadvantage of hybridization in these regions, or of the short localization scale. The same basic signal is seen for the 200-member experiments introduced below.

We conclude that the localization changes have a neutral impact overall, and most of the further results will use the new settings. However, the mixed results in different regions of the atmosphere suggest that there could be benefit in using different localization settings in different regions. Such a modification has recently been implemented by Environment and Climate Change Canada (Houtekamer *et al.*, 2014).

4.2. *Varying the ensemble error modes and weight*

We now consider the effect of changing the ensemble data used as input to hybrid 4D-Var. Given the substantial changes in the performance of the ensemble reported in Bowler *et al.* (2017), do we see a consequent change in the DA when these data are used?

As a first test of this, we compare two experiments, which both give a high weight ($\beta_c^2 = 1.0$) to the climatological covariance as in our current operational system and a weight of $\beta_e^2 = 0.3$ to the ensemble covariance. We then compare the performance of a DA system using ETKF and En-4DEnVar ensembles (the latter being the experiment 20 described in Bowler *et al.* (2017) using relaxation to prior perturbations (RTPP) with a scaling of 0.5 and relaxation to prior spread (RTPS) with a scaling of 0.9). A verification scorecard for this comparison is shown in Figure 3. For most of the variables and lead times, the change to using En-4DEnVar error modes provides a small benefit. Since a small weight is given to the ensemble covariances, one would not expect the impact to be large. Nonetheless, it is significant that the impact can be seen across a wide range of variables. More detail on the changes to the ensemble, including how the covariances are affected, is contained in Bowler *et al.* (2017).

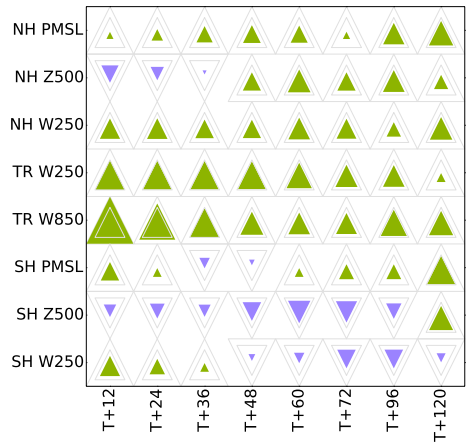

Figure 4. As Figure 1, but showing change from using En-4DEnVar rather than ETKF error modes when a high weight is given to the ensemble covariance (i.e. comparing 4DVar.envar44.30c70e.NewLoc with 4DVar.etkf44.30c70e.NewLoc). [Colour figure can be viewed at wileyonlinelibrary.com].

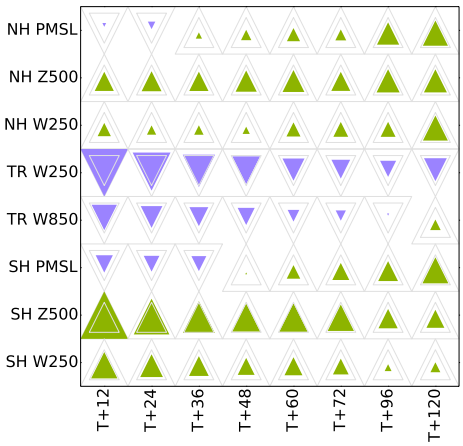

Figure 6. Relative change in RMSE for NWP index fields of changing the hybrid weights from $\beta_c^2 = 1$ and $\beta_e^2 = 0.3$ to $\beta_c^2 = 0.3$ and $\beta_e^2 = 0.7$ while using En-4DEnVar error modes (i.e. comparing 4DVar.envar44.30c70e.NewLoc with 4DVar.envar44.100c30e.NewLoc). [Colour figure can be viewed at wileyonlinelibrary.com].

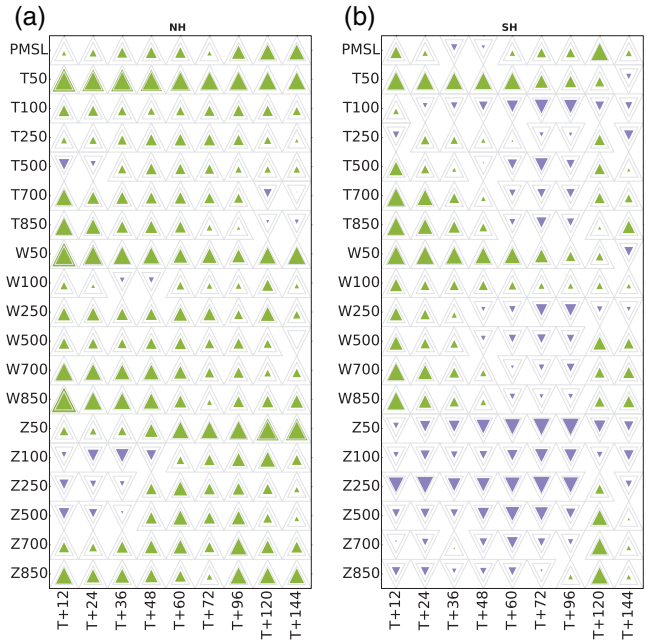

Figure 5. Impact of using En-4DEnVar rather than ETKF error modes when a high weight is given to the ensemble covariance. Extended verification information for the (a) Northern and (b) Southern Extratropics (i.e. comparing 4DVar.envar44.30c70e.NewLoc with 4DVar.etkf44.30c70e.NewLoc). [Colour figure can be viewed at wileyonlinelibrary.com].

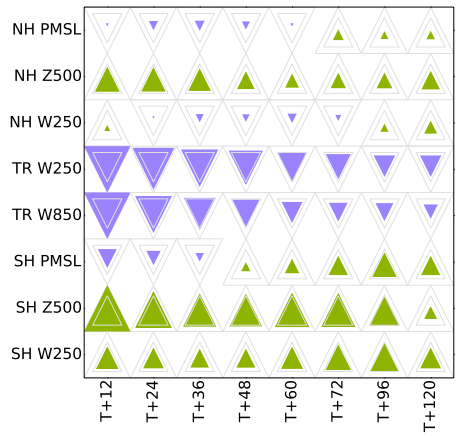

Figure 7. Relative change in RMSE for NWP index fields of changing the hybrid weights from $\beta_c^2 = 1$ and $\beta_e^2 = 0.3$ to $\beta_c^2 = 0.3$ and $\beta_e^2 = 0.7$ while using ETKF error modes (i.e. comparing 4DVar.etkf44.30c70e.NewLoc with 4DVar.etkf44.100c30e.NewLoc). [Colour figure can be viewed at wileyonlinelibrary.com].

Other authors (Todling and El Akkraoui, 2014; Buehner *et al.*, 2016) have seen only small differences in the performance of hybrid DA schemes when changing the source of the ensemble covariances. Here we have noted substantial improvements. This is probably due to the fact that there are substantial differences between the ETKF ensemble and the 4DEnVar ensemble. The results comparing a non-hybrid 4D-Var (Figure 1) show that there are clear benefits from using a hybrid 4D-Var with the ETKF ensemble. Thus we conclude that the Met Office ETKF, despite its limitations, provides useful covariances to the DA, but that the 4DEnVar ensemble provides even better covariances than this.

Our next set of tests uses $\beta_c^2 = 0.3$ and $\beta_e^2 = 0.7$. With a high weight given to the ensemble covariance, we expect a larger impact from changing the ensemble input to the DA. Figure 4 shows the impact of changing from ETKF error modes to En-4DEnVar error modes with these weights. As expected, the impacts are larger than for the situation where a high weight is given to the climatological covariance. The verification results are mostly positive, with the notable exception of 500 hPa geopotential height in the southern Extratropics. Looking at verification information for more variables (Figure 5), we see improvement for most variables in the northern Extratropics. In the southern Extratropics, we see benefit for many variables apart from geopotential height, which often has poorer accuracy. Overall this corresponds to a 1.1% change in the Met Office's NWP index.

An interesting perspective on this result is obtained by comparing the effect of increasing the weight given to the ensemble covariance in experiments using En-4DEnVar and ETKF error modes. Figure 6 shows the impact with En-4DEnVar error modes. The forecasts become substantially worse in the Tropics, but much better in the Extratropics, particularly for 500 hPa geopotential height. Overall the change in the Met Office's NWP index is small (less than 0.1%), but this masks some large changes in its components. The corresponding results using ETKF error modes are shown in Figure 7. The change in extratropical performance is similar, but the degradation in the Tropics is worse, leading to a worse performance overall.

Given the above results, we conclude that the 4DEnVar ensemble better represents the forecast errors. However, we also note that the 4DEnVar ensemble is underspread for most variables (Bowler *et al.*, 2017), and perhaps it is this change in spread which is leading to the improved performance. To investigate this, we artificially increased the weight given to the ensemble covariance, as this has the same effect as post-processing the ensemble to increase its spread. The ratio of the spread of the ETKF ensemble to that of the 4DEnVar ensemble was calculated

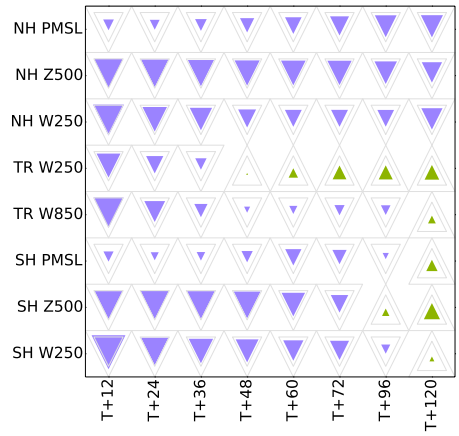

Figure 8. Relative change in RMSE for NWP index fields of giving extra weight to the ensemble covariance while using En-4DEnVar error modes. This is testing the effect of overweighting the ensemble to compensate for a lack of spread (i.e. comparing 4DVar.envar44.30c140e.NewLoc with 4DVar.envar44.30c70e.NewLoc). [Colour figure can be viewed at wileyonlinelibrary.com].

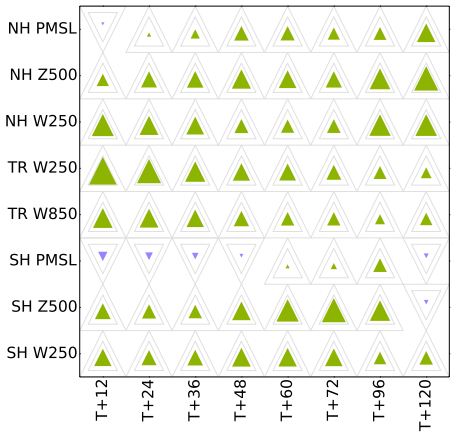

Figure 9. Relative change in RMSE for NWP index fields of increasing the ensemble size from 44 to 200 members (i.e. comparing 4DVar.envar200.100c30e.NewLoc with 4DVar.envar44.100c30e.NewLoc). Both experiments use En-4DEnVar error modes and a high weighting to the climatological covariance ($\beta_c^2 = 1$ and $\beta_e^2 = 0.3$). [Colour figure can be viewed at wileyonlinelibrary.com].

for a number of different regions and variables. For mean-sea-level pressure in the Northern Hemisphere, the spread of the ETKF ensemble is over twice that of the 4DEnVar ensemble. By contrast, the spread of the 4DEnVar ensemble is greater than that of the ETKF ensemble for wind speed at 250 hPa for all regions considered. The reason for this is that the analysis increments used in the additive inflation (Bowler *et al.*, 2017) have a relatively larger spread in wind compared to the potential temperature and pressure, which is itself a symptom of limitations in the Met Office's current climatological covariance model (Wlasak and Cullen, 2014). Aside from these outliers, there was a fairly consistent signal that the ETKF spread is around 40% greater than the spread of the 4DEnVar ensemble. Therefore, we ran a test with ensemble weighting $\beta_e^2 = 1.4$, which is twice that used in the conventional experiments ($\beta_c^2 = 0.3$ was retained). We compare the performance of this experiment with the previous test which used $\beta_e^2 = 0.7$ (Figure 8). As can be seen, overweighting the ensemble covariance to compensate for a lack of spread gives a largely negative impact. This could be due to the fact that increasing the weight in this way amplifies the mid- and upper-level wind increments too much since these variables are are not so underspread as other variables.

4.3. *Increasing the ensemble size*

Given the apparent benefit of changing from ETKF error modes to En-4DEnVar error modes we now ask whether there is additional benefit from increasing the ensemble size. In Bowler *et al.* (2017) we discuss the set-up of a 200 member 4DEnVar ensemble. This ensemble has a slightly smaller spread than the equivalent 44 member ensemble. Nonetheless, the increased ensemble size should mean that it is able to produce a better analysis than the smaller ensemble. Additionally the increased ensemble size means that it is justifiable to decrease the severity of the localization used in the DA. For the 200 member ensemble we have chosen to increase the horizontal localization distance of 600 km (approximately equivalent to a Gaspari–Cohn localization cut-off distance of 2190 km) to 800 km (equivalent cut-off distance of 2920 km) and from 1.5 to 2 scale-heights in the vertical (cf. the 44 member ensemble settings in section 4.1). The rationale for this increase is the same as that of Bowler *et al.* (2017), namely that the increase in the observation volume permitted to influence a grid-point is proportionally less than the increase in the ensemble size.

Figure 9 shows the impact of increasing the ensemble size when a high weight is given to the climatological covariance (that is $\beta_c^2 = 1$ and $\beta_e^2 = 0.3$). The results indicate a clear benefit to the performance of hybrid 4D-Var (a 0.5% change in the NWP index). A similar generally positive effect, but with a smaller

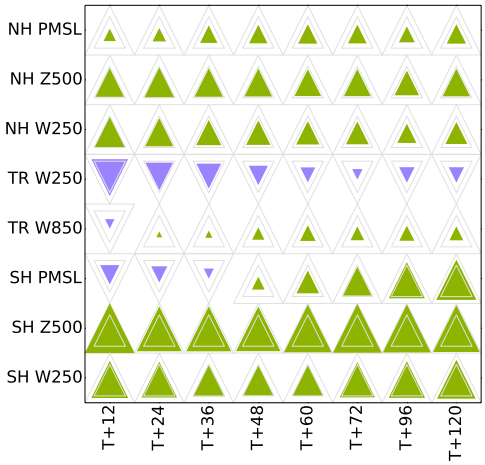

Figure 10. Relative change in RMSE for NWP index fields of changing the covariance weightings from $\beta_c^2 = 1.0$ and $\beta_e^2 = 0.3$ to $\beta_c^2 = 0.3$ and $\beta_e^2 = 0.7$ when using 200 En-4DEnVar error modes (i.e. comparing 4DVar.envar200.30c70e.NewLoc with 4DVar.envar200.100c30e.NewLoc). [Colour figure can be viewed at wileyonlinelibrary.com].

average magnitude, was seen in the earlier set of experiments for the same covariance weightings when we increased the (ETKF-based) ensemble size from 23 to 176 members, though in those experiments we did not alter the localization length-scales to account for the increased ensemble size.

Figure 10 shows the additional impact when we increase the weight given to the 200 member ensemble covariance. As with Figure 6 – the corresponding figure for 44-member experiments – there are degradations in tropical performance, but improvements in extratropical performance. With the large ensemble size the benefits are now very large, and the NWP index score has improved overall, increasing by around 0.8%. In contrast, with the earlier 176 member ETKF-based experiments (not shown) we saw a degradation in performance when increasing the weight given to the ensemble covariance. This provides further evidence that En-4DEnVar is providing a better characterization of short-period forecast errors than the ETKF.

Now we consider the combined effect of the increased ensemble size, localization and hybrid weighting changes with the improvement from using En-4DEnVar. Comparing a hybrid 4D-Var trial using all these changes with a trial using the operational settings (including the use of the ETKF), we see large improvements (Figure 11). This gives a 1.5% change in the NWP index and demonstrates that large benefits are available within the context of hybrid 4D-Var. The more comprehensive verification results shown in Figure 12 show consistent forecast improvements in nearly all regions of the globe.

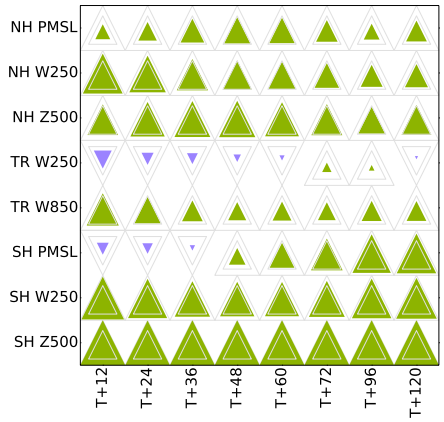

Figure 11. Relative change in RMSE for NWP index fields of changing the ensemble generation technique (to En-4DEnVar from ETKF), the ensemble size (to 200 from 44), the covariance weightings (to $\beta_c^2 = 0.3$ and $\beta_e^2 = 0.7$ from $\beta_c^2 = 1$ and $\beta_e^2 = 0.3$), and using the tighter localization settings (i.e. comparing 4DVar.envar200.30c70e.NewLoc with 4DVar.etkf44.100c30e.OpLoc). [Colour figure can be viewed at wileyonlinelibrary.com].

5. Comparing hybrid 4D-Var and hybrid 4DEnVar

In the previous section we have examined the effect of changes to the ensemble used as input to hybrid 4D-Var. This focused on results from the second trial period. Now we consider the comparison between hybrid 4D-Var and hybrid 4DEnVar and whether this is affected by the ensemble changes.

In experiments for the first trial period, we find that hybrid 4DEnVar does not perform as well as hybrid 4D-Var, similar to the results of Lorenc *et al.* (2015). When using the larger 176 member ensemble, hybrid 4DEnVar performs best when the ensemble covariances are given a high weight, but hybrid 4D-Var performs marginally better when a high weight is given to the static covariances. With a high weight to the ensemble the comparison with hybrid 4D-Var is as shown in Figure 13. Note the switch to verification against observations rather than ECMWF analyses since ECMWF analysis verification was not available for this earlier trial. Also, winds are verified against sondes, rather than satellites and aircraft as in the later graphs. This confirms that hybrid 4DEnVar performs worse than hybrid 4D-Var for most variables.

Figure 14 shows the comparison between hybrid 4D-Var and hybrid 4DEnVar when using error modes from the 44 member 4DEnVar ensemble. This and following graphs use verification against observations to be comparable with the results in Figure 13. It is clear that hybrid 4DEnVar is worse than hybrid 4D-Var across a wide range of variables and the detriment is often larger than for the previous comparison which used an ETKF-based ensemble. Thus changing the source of the ensemble data can make substantial changes to the performance of the DA, but these improvements do not necessarily benefit hybrid 4DEnVar more than hybrid 4D-Var.

Next we consider the comparison between hybrid 4DEnVar and hybrid 4D-Var when using error modes from the 200 member 4DEnVar ensemble (Figure 15). Again, hybrid 4DEnVar performs substantially worse than hybrid 4D-Var, although the difference is marginally smaller with the large ensemble size, especially at short lead times.

5.1. *Removing the climatological covariance*

Given that the presence of a climatological covariance makes a large impact on the comparison between hybrid 4D-Var and hybrid 4DEnVar, it is reasonable to run a test using only ensemble covariances. Previously we had not been able to run such a test because the ETKF ensemble does not provide sufficiently high-quality information for the forecast/assimilation system to be stable. Even with the removal of the climatological covariance

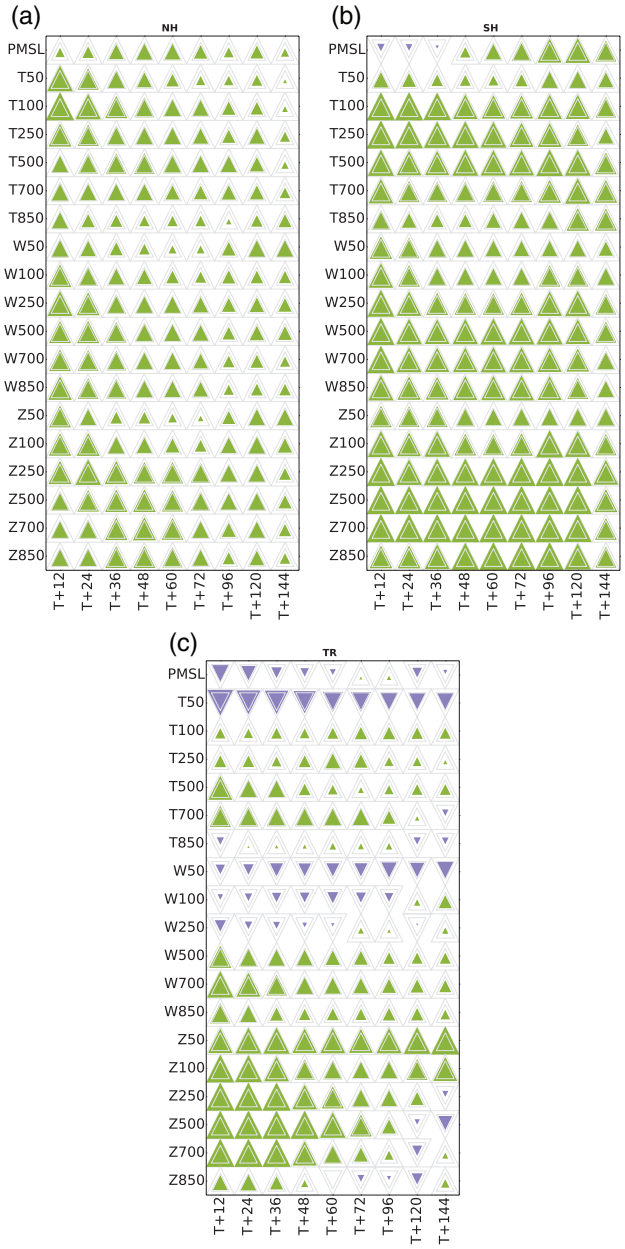

Figure 12. As Figure 11, but for a more comprehensive set of fields (i.e. comparing 4DVar.envar200.30c70e.NewLoc with 4DVar.etkf44.100c30e.OpLoc). The verification areas are (a) Northern Extratropics, (b) Southern Extratropics and (c) Tropics. [Colour figure can be viewed at wileyonlinelibrary.com].

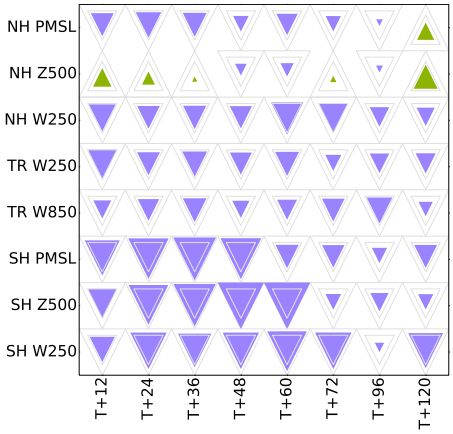

Figure 13. Relative change in RMSE for NWP index fields of changing to hybrid 4DEnVar from hybrid 4D-Var. Both experiments use input from a 176 member ETKF ensemble, giving a high weighting ($\beta_c^2 = 0.3$ and $\beta_e^2 = 0.7$) to the ensemble covariance (i.e. comparing 4DEnVar.etkf176.30c70e.OpLoc with 4DVar.etkf176.30c70e.OpLoc). Verification is against observations from 0000 UTC on 8 August 2012 to 31 August 2012. [Colour figure can be viewed at wileyonlinelibrary.com].

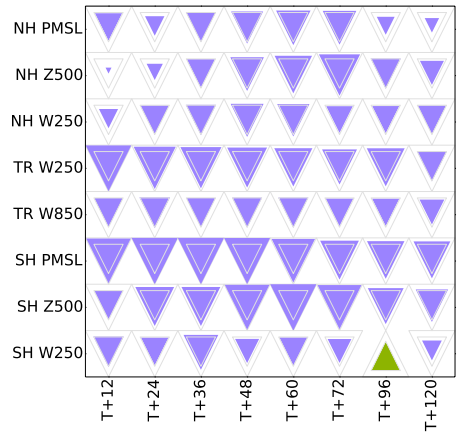

Figure 14. As Figure 13, but for 44 member ensembles of 4DEnVars (i.e. comparing 4DEnVar.envar44.30c70e.NewLoc with 4DVar.envar44.30c70e.NewLoc). Verification is against observations from 0000 UTC on 31 January 2014 to 0000 UTC on 19 March 2014. [Colour figure can be viewed at wileyonlinelibrary.com].

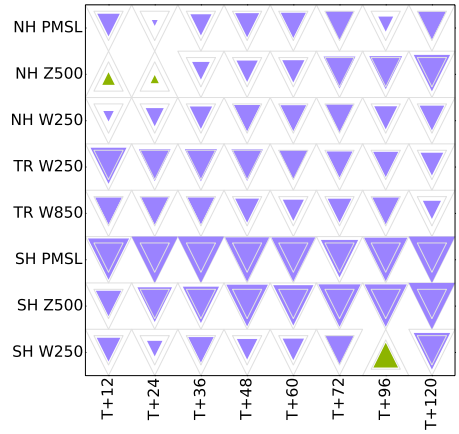

Figure 15. As Figure 14, but for 200 member ensembles of 4DEnVars and giving the same high weighting to the ensemble covariance (i.e. comparing 4DEnVar.envar200.30c70e.NewLoc with 4DVar.envar200.30c70e.NewLoc). [Colour figure can be viewed at wileyonlinelibrary.com].

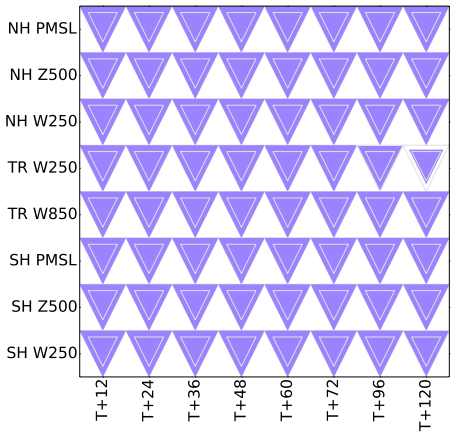

Figure 16. As Figure 14, but for both experiments using the ensemble covariance alone with $\beta_e^2 = 1$ and $\beta_c^2 = 0$ (i.e. comparing 4DEnVar.envar44.0c100e.NewLoc with 4DVar.envar44.0c100e.NewLoc). [Colour figure can be viewed at wileyonlinelibrary.com].

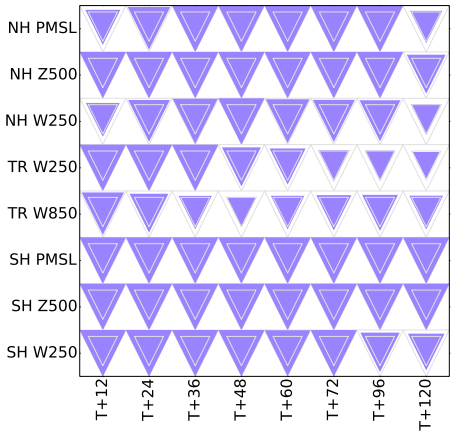

Figure 17. As Figure 16, but with both the experiments using large localization length-scales (i.e. comparing 4DEnVar.envar44.0c100e.BigLoc with 4DVar.envar44.0c100e.BigLoc). [Colour figure can be viewed at wileyonlinelibrary.com].

there are still a few remaining differences between 4D-Var and 4DEnVar. 4D-Var propagates the covariance information using an explicitly coded linear model, whereas 4DEnVar uses an ensemble of nonlinear trajectories. Furthermore, 4D-Var applies localization at the start of the DA window, and propagates its effect through the window using the linear model whereas 4DEnVar ignores differences in time when localising 4D covariances.

Figure 16 shows the comparison between 4D-Var and 4DEnVar when using the ensemble covariance alone with $\beta_e^2 = 1$ and $\beta_c^2 = 0$. The difference is much larger than when using a hybrid background-error covariance matrix. This is a puzzling result, since we assumed that the treatment of the climatological covariance was the main difference between the two DA methods, and the gap in performance is larger in the hybrid when a high weight is given to the ensemble covariance. One possible explanation is that this is being caused by our short localization length-scales. Idealized studies (Fairbairn *et al.*, 2013; Poterjoy and Zhang, 2015; Bocquet, 2016) show that not propagating the localization in time puts 4DEnVar at a disadvantage compared to hybrid 4D-Var, which only applies the localization at the initial time, implicitly propagating the localized covariance in time. Lorenc *et al.* (2015) presented tests with single observations in situations where the effect should be important – a jet stream and a tropical cyclone. With a 1200 km localization length-scale, their 4DEnVar could perform as well as 4D-Var with $\beta_e^2 = 1$ and $\beta_c^2 = 0$. We have run a test using increased localization length-scales in both 4D-Var and 4DEnVar – from 600 to 1200 km in the horizontal and from 1.5 to 2 scale-heights in the vertical. Figure 17 shows the comparison between 4DEnVar and 4D-Var using $\beta_e^2 = 1$ and $\beta_c^2 = 0$ with the longer localization length-scales. Although the difference is reduced, 4DEnVar is still much worse than 4D-Var. Further tests are needed to fully understand what are the important factors in the difference between 4D-Var and 4DEnVar. Nevertheless our results do show that, for the shorter length-scales which give best results with our ensembles, the time-independent localization in our 4DEnVar is detrimental. Methods to address this by advecting the localization have been proposed by Bocquet (2016), Desroziers *et al.* (2016) and Frolov and Bishop (2016).

6. Results summary

The Met Office's NWP index provides a single number which can be used to summarize the overall performance of a system. Therefore, we use this index as a way to summarize the main results in this article. Figure 18 shows the main results for both sets of experiments when the verification is performed against observations. The hybrid 4D-Var run which uses ETKF error modes with operational localization settings and covariance weightings is considered the baseline, and all the other results are plotted in comparison to this. The main features we see (focusing mainly on Figure 18(b)) are:

- hybrid 4DEnVar using covariances from a 200 member En-4DEnVar performs better than the non-hybrid 4D-Var which uses only climatological covariances (purple dashed lines and black square);
- hybrid 4D-Var using a 200 member En-4DEnVar performs much better than the current operational system, around

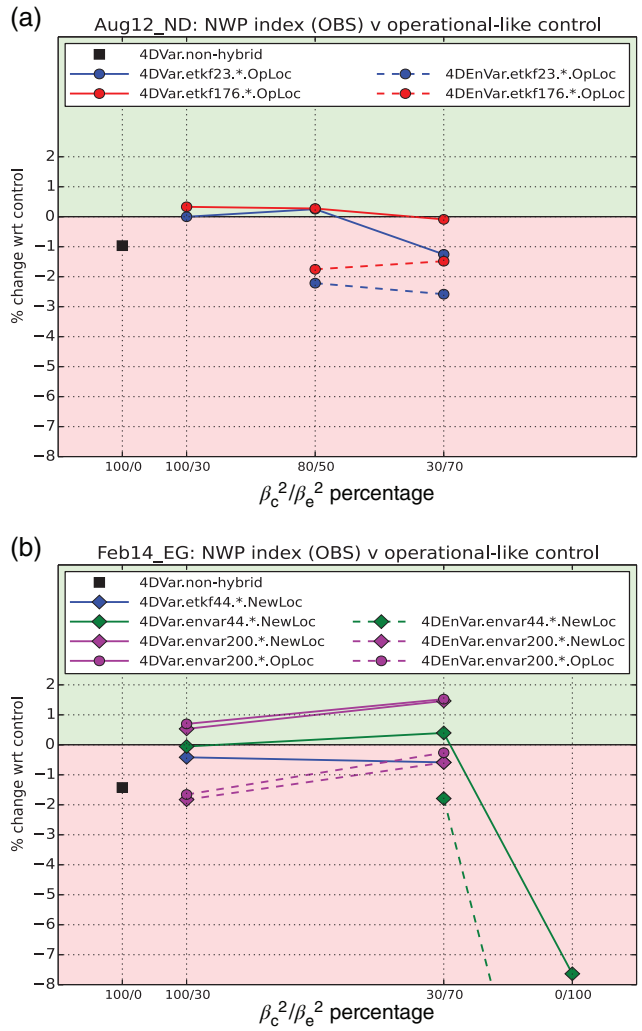

Figure 18. Summary of the changes in the NWP index for the core experiments from the (a) first and (b) second set of experiments. (Tables 1 and 2 define the experiment IDs in the legend.)

2% better than any hybrid 4DEnVar system tested and around 3% better than a non-hybrid 4D-Var (purple and blue solid lines and black square);

- there are small benefits from moving to En-4DEnVar whilst keeping the ensemble size at 44 members and these benefits are greater when increasing the weight to the ensemble covariance (blue and green solid lines);
- both hybrid 4D-Var and hybrid 4DEnVar benefit from increasing the weight given to the ensemble covariance when using the 200 member 4DEnVar ensemble (purple solid and dashed lines).

7. Conclusions

In this work we have investigated changes to the DA system – changing the source of the ensemble input data, the ensemble size, the weight to the ensemble covariance, comparing 4DEnVar with 4D-Var and changing the localization settings.

When changing from using an ETKF-based ensemble to using an En-4DEnVar-based ensemble, a small benefit is seen for hybrid 4D-Var if the background-error covariance is weighted highly towards the static (climatological) covariance matrix. When a high weight is given to the ensemble-based covariance, the improvement is larger with greatest benefits being seen in the Tropics (Figures 4 and 5). In this case the benefit of changing the ensemble input data is seen across a wide range of variables in the Northern Hemisphere. Results in the Southern Hemisphere are more mixed and are mostly negative for geopotential heights.

The effect of changing the weight given to the ensemble covariance is broadly similar for both ensembles. The forecasts are generally better in the Extratropics for the higher weight given to the ensemble covariance, and worse in the Tropics (Figures 6 and 7). For the 4DEnVar ensemble, the forecast detriment is generally smaller in the Tropics than with the ETKF, but the benefits are generally similar in the Extratropics. This means there is an opportunity to further increase the benefits from using hybrid covariances by addressing the specific issues arising in the Tropics, possibly by using more sophisticated localization methods, such as waveband localization (Buehner, 2012).

Tests using a 200 member En-4DEnVar have shown that increasing the ensemble size is beneficial for almost every forecast variable (Figure 9). It is quite unusual to see a benefit across such a wide variety of variables and regions. When an ETKF-based ensemble is used to provide the ensemble covariance information, there is a small benefit to the DA from increasing the ensemble size from 23 to 176 members (Figure 18). Giving greater weight to the ETKF ensemble covariance is detrimental to the performance of hybrid 4D-Var. However, giving greater weight to the ensemble covariance is beneficial to hybrid 4DEnVar.

We have run a small number of experiments using the ensemble covariance alone. As expected, there is a large degradation in 4D-Var performance when the climatological covariance is removed, but the degradation in 4DEnVar is larger (Figure 16), and much worse than expected. Experiments with the localization length-scales indicate that this is largely due to 4DEnVar using localization length-scales which are too short (Figure 17). We hope to conduct further experiments in this area to further understand the causes of the poor performance of 4DEnVar.

Given the good performance of hybrid 4D-Var when using a 4DEnVar ensemble, En-4DEnVar is an attractive alternative ensemble prediction system to the operational ETKF. As noted in Bowler *et al.* (2017), there are a number of issues to be addressed before this can be done. In addition we note substantial degradations in tropical performance when giving a high weight to the ensemble covariance (Figure 6), which will need to be addressed. By contrast, hybrid 4DEnVar is not in a position to replace hybrid 4D-Var as the Met Office's operational DA scheme. However, this is largely because the focus of our work so far has been on generating the ensemble and less effort has been applied to exploiting it in the DA step itself. It seems that a major issue is the way time-localization is performed in our hybrid 4DEnVar – the localization region is not propagated. Bocquet (2016) and Frolov and Bishop (2016) have discussed this and Desroziers *et al.* (2016) have presented a practical method of advecting the localization in 4DEnVar, perhaps using velocities slightly less than the smoothed wind field.

At the start of this work, we hoped to find a combination of ensemble size, weighting and localization length-scales which reduced the performance gap of hybrid 4D-Var over hybrid 4DEnVar seen by Lorenc *et al.* (2015). The combinations we tested did not achieve this – improvements to the ensemble and localization benefited hybrid 4D-Var as well as hybrid 4DEnVar. One of the combinations tested did achieve a similar performance to the operational configuration – hybrid 4DEnVar giving a high weight to the 200 member En-4DEnVar and using operational localization settings. However, this was still substantially worse than a hybrid 4D-Var assimilation with the same settings. In the longer-term, computational challenges exist with the use of 4D-Var, and so an improved 4DEnVar remains an aspiration for operational use beyond 2020.

Acknowledgements

We are grateful to the Korea Meteorological Administration for hosting Adam Clayton and contributing supercomputing resources for this project. We thank Stephen Oxley and Brett Candy at the Met Office for their crucial and timely support and the reviewers for their helpful comments.

Appendix

Met Office NWP Index

The Met Office's NWP index provides a single number which may be used to summarize the performance of a forecasting system. Let $R^e_{v,t}$ be the RMS error of the forecasts from experiment e for variable v and forecast lead time t. The NWP experiments are compared with a persistence forecast, calculated by verifying the analysis from t hours ago against the current observations/analysis. Using the persistence forecast allows the calculation of a skill score for the NWP forecast as

$$S^e_{v,t} = 1 - \left(\frac{R^e_{v,t}}{R^p_{v,t}}\right)^2, \tag{A1}$$

where $R^p_{v,t}$ is the RMS error of the persistence forecast. Note that, because the persistence forecast is based on the analyses from the experiment being verified, it is not an identical reference between forecasts. However, since errors in a persistence forecast are often large, it is not expected that changes in the analysis system will have a substantial effect on this quantity.

The variables for which the skill score is calculated are chosen to be those which are important to Met Office customers. The variables are pressure at mean sea level, 500 hPa geopotential height and wind vector at 250 hPa in the Northern and Southern Extratropics (>20° latitude north or south). In the Tropics, forecasts of wind vector at 250 and 850 hPa are assessed. The forecasts initialized at 0000 UTC and 1200 UTC each day are verified at lead times of T+12, 24, 36, 48, 60, 72, 96 and 120 h. To compose the index, the various skill scores are combined using weights assigned to the individual elements according to

$$I^e = \frac{1}{100}\sum_{v,t} w_{v,t} S^e_{v,t}, \tag{A2}$$

where the weights for each of the variables and lead times are given in Table A1.

For many of the experiments, we report verification which has been performed against ECMWF analyses. For these both the forecasts and analyses have been interpolated to a 1.5° regular latitude–longitude grid before comparison. For other experiments we report the verification measured against observations. Bilinear interpolation is used to calculate forecast values at the observation positions.

Table A1. Percentage weights given to the different components of the NWP index for short-range (T + 12, T + 24, T + 36, T + 48, T + 60, T + 72) and long-range (T + 96, T + 120) forecasts.

Variable	Short-range	Long-range
NH PMSL	3.2	6.4
NH H500	1.2	2.4
NH W250	1.2	2.4
TR W850	1.0	2.0
TR W250	0.6	1.2
SH PMSL	1.6	3.2
SH H500	0.6	1.2
SH W250	0.6	1.2

References

Bishop CH, Etherton BJ, Majumdar SJ. 2001. Adaptive sampling with the ensemble transform Kalman filter. Part 1: Theoretical aspects. *Mon. Weather Rev.* **129**: 420–436, doi: 10.1175/1520-0493(2001)129<0420:ASWTET>2.0.CO;2.

Bocquet M. 2016. Localization and the iterative ensemble Kalman smoother. *Q. J. R. Meteorol. Soc.* **142**: 1075–1089, doi: 10.1002/qj.271110.1002/qj.2711.

Bowler NE, Arribas A, Mylne KR, Robertson KB, Beare SE. 2008. The MOGREPS short-range ensemble prediction system. *Q. J. R. Meteorol. Soc.* **134**: 703–722, doi: 10.1002/qj.23410.1002/qj.234.

Bowler NE, Clayton AM, Jardak M, Lee E, Lorenc AC, Piccolo C, Pring SR, Wlasak MA, Barker DM, Inverarity GW, Swinbank R. 2017. Inflation and localization tests in the development of an ensemble of 4D-ensemble variational assimilations. *Q. J. R. Meteorol. Soc.*, doi: 10.1002/qj.3004.

Buehner M. 2012. Evaluation of a spatial/spectral covariance localization approach for atmospheric data assimilation. *Mon. Weather Rev.* **140**: 617–636, doi: 10.1175/MWR-D-10-05052.1.

Buehner M, Houtekamer PL, Charette C, Mitchell HL, He B. 2010a. Intercomparison of variational data assimilation and the ensemble Kalman filter for global deterministic NWP. Part I: Description and single-observation experiments. *Mon. Weather Rev.* **138**: 1550–1566, doi: 10.1175/2009MWR3157.1.

Buehner M, Houtekamer PL, Charette C, Mitchell HL, He B. 2010b. Intercomparison of variational data assimilation and the ensemble Kalman filter for global deterministic NWP. Part II: One-month experiments with real observations. *Mon. Weather Rev.* **138**: 1567–1586, doi: 10.1175/2009MWR3158.1.

Buehner M, McTaggart-Cowan R, Beaulne A, Charette C, Garand L, Heilliette S, Lapalme E, Laroche S, Macpherson SR, Morneau J, Zadra A. 2015. Implementation of deterministic weather forecasting systems based on ensemble-variational data assimilation at Environment Canada. Part I: The global system. *Mon. Weather Rev.* **143**: 2532–2559, doi: 10.1175/MWR-D-14-00354.1.

Buehner M, McTaggart-Cowan R, Heilliette S. 2016. An ensemble Kalman filter for numerical weather prediction based on variational data assimilation: VarEnKF. *Mon. Weather Rev.*, doi: 10.1175/MWR-D-16-0106.1.

Candy B. 2014. 'Assimilation of satellite data for the land surface'. In *Annual Seminar on Use of Satellite Observations in Numerical Weather Prediction. ECMWF: Reading, UK.*

Clayton AM, Lorenc AC, Barker DM. 2013. Operational implementation of a hybrid ensemble/4D-Var global data assimilation system at the Met Office. *Q. J. R. Meteorol. Soc.* **139**: 1445–1461, doi: 10.1002/qj.2054.

Davies T, Cullen MJP, Malcolm A, Mawson M, Staniforth A, White AA, Wood N. 2005. A new dynamical core for the Met Office's global and regional modelling of the atmosphere. *Q. J. R. Meteorol. Soc.* **131**: 1759–1782, doi: 10.1256/qj.04.101.

Desroziers G, Arbogast E, Berre L. 2016. Improving spatial localization in 4DEnVar. *Q. J. R. Meteorol. Soc.* **142**: 3171–3185, doi: 10.1002/qj.2898.

Dharssi I, Bovis KJ, Macpherson B, Jones CP. 2011. Operational assimilation of ASCAT surface soil wetness at the Met Office. *Hydrol. Earth Syst. Sci.* **15**: 2729–2746, doi: 10.5194/hess-15-2729-2011.

Donlon CJ, Martin M, Stark J, Roberts-Jones J, Fiedler E, Wimmer W. 2012. The Operational Sea Surface Temperature and Sea Ice Analysis (OSTIA) system. *Remote Sens. Environ.* **116**: 140–158, doi: 10.1016/j.rse.2010.10.017.

Fairbairn D, Pring SR, Lorenc AC, Roulstone I. 2013. A comparison of 4D-Var with ensemble data assimilation methods. *Q. J. R. Meteorol. Soc.* **140**: 281–294, doi: 10.1002/qj.2135.

Fisher M. 2003. 'Background-error covariance modelling'. In Proceedings of Seminar on Recent Developments in Data Assimilation for Atmosphere and Ocean, 8–12 September. ECMWF: Reading, UK, pp. 45–64.

Frolov S, Bishop CH. 2016. Localized ensemble-based tangent linear models and their use in propagating hybrid error covariance models. *Mon. Weather Rev.* **144**: 1383–1405, doi: 10.1175/MWR-D-15-0130.1.

Gaspari G, Cohn SE. 1999. Construction of correlation functions in two and three dimensions. *Q. J. R. Meteorol. Soc.* **125**: 723–757, doi: 10.1002/qj.49712555417.

Gauthier P, Thépaut J-N. 2001. Impact of the digital filter as a weak constraint in the preoperational 4D-VAR assimilation system of Météo-France. *Mon. Weather Rev.* **129**: 2089–2102, doi: 10.1175/1520-0493(2001)129<2089:IOTDFA>2.0.CO;2.

Hamill TM, Snyder C. 2000. A hybrid ensemble Kalman filter–3D variational analysis scheme. *Mon. Weather Rev.* **128**: 2905–2919.

Houtekamer PL, Deng X, Mitchell HL, Baek S-J, Gagnon N. 2014. Higher resolution in an operational ensemble Kalman filter. *Mon. Weather Rev.* **142**: 1143–1162, doi: 10.1175/MWR-D-13-00138.1.

Kepert JD. 2009. Covariance localization and balance in an ensemble Kalman filter. *Q. J. R. Meteorol. Soc.* **135**: 1157–1176, doi: 10.1002/qj.443.

Kleist DT, Ide K. 2015a. An OSSE-based evaluation of hybrid variational-ensemble data assimilation for the NCEP GFS. Part I: System description and 3D-hybrid results. *Mon. Weather Rev.* **143**: 433–451, doi: 10.1175/MWR-D-13-00351.1.

Kleist DT, Ide K. 2015b. An OSSE-based evaluation of hybrid variational-ensemble data assimilation for the NCEP GFS. Part II: 4DEnVar and hybrid variants. *Mon. Weather Rev.* **143**: 452–470, doi: 10.1175/MWR-D-13-00350.1.

Liu C, Xiao Q, Wang B. 2008. An ensemble-based four-dimensional variational data assimilation scheme. Part I: Technical formulation and preliminary test. *Mon. Weather Rev.* **136**: 3363–3373, doi: 10.1175/2008MWR2312.1.

Lorenc AC, Rawlins F. 2005. Why does 4D-Var beat 3D-Var? *Q. J. R. Meteorol. Soc.* **131**: 3247–3257, doi: 10.1256/qj.05.85.

Lorenc AC, Ballard SP, Bell RS, Ingleby NB, Andrews PLF, Barker DM, Bray JR, Clayton AM, Dalby T, Li D, Payne TJ, Saunders FW. 2000. The Met Office global three-dimensional variational data assimilation scheme. *Q. J. R. Meteorol. Soc.* **126**: 2991–3012, doi: 10.1002/qj.49712657002.

Lorenc AC, Bowler NE, Clayton AM, Pring SR, Fairbairn D. 2015. Comparison of hybrid-4DEnVar and hybrid-4DVar data assimilation methods for global NWP. *Mon. Weather Rev.* **143**: 212–229, doi: 10.1175/MWR-D-14-00195.1.

Poterjoy J, Zhang F. 2015. Systematic comparison of four-dimensional data assimilation methods with and without the tangent linear model using hybrid background-error covariance: E4DVar versus 4DEnVar. *Mon. Weather Rev.* **143**: 1601–1621, doi: 10.1175/MWR-D-14-00224.1.

Poterjoy J, Zhang F. 2016. Comparison of hybrid four-dimensional data assimilation methods with and without the tangent linear and adjoint models for predicting the life cycle of hurricane *Karl* (2010). *Mon. Weather Rev.* **144**: 1449–1468, doi: 10.1175/MWR-D-15-0116.1.

Rabier F. 2005. Overview of global data assimilation developments in numerical weather-prediction centres. *Q. J. R. Meteorol. Soc.* **131**: 3215–3233, doi: 10.1256/qj.05.129.

Rawlins F, Ballard SP, Bovis KJ, Clayton AM, Li D, Inverarity GW, Lorenc AC, Payne TJ. 2007. The Met Office global four-dimensional variational data assimilation scheme. *Q. J. R. Meteorol. Soc.* **133**: 347–362, doi: 10.1002/qj.32.

Schmetz J, Holmlund K, Hoffman J, Strauss B, Mason B, Gaertner V, Koch A, Vandeberg L. 1993. Operational cloud-motion winds from Meteosat infrared images. *J. Appl. Meteorol.* **32**: 1206–1225, doi: 10.1175/1520-0450(1993)032<1206:OCMWFM>2.0.CO;2.

Talagrand O. 2010. Variational assimilation. In *Data Assimilation: Making Sense of Observations*, Lahoz W, Khattatov B, Ménard R. (eds.): 41–67. Springer: Berlin.

Todling R, El Akkraoui A. 2014. *Hybrid Data Assimilation without Ensemble Filtering*. Preprint 20140011180, NASA Goddard Space Flight Center: Greenbelt, MD. http://ntrs.nasa.gov/search.jsp?R=20140011180 (accessed 10 December 2016).

Walters DN, Best MJ, Bushell AC, Copsey D, Edwards JM, Falloon PD, Harris CM, Lock AP, Manners JC, Morcrette CJ, Roberts MJ, Stratton RA, Webster S, Wilkinson JM, Willett MR, Boutle IA, Earnshaw PD, Hill PG, MacLachlan C, Martin GM, Moufouma-Okia W, Palmer MD, Petch JC, Rooney GG, Scaife AA, Williams KD. 2011. The Met Office Unified Model global atmosphere 3.0/3.1 and JULES Global Land 3.0/3.1 configurations. *Geosci. Model Dev.* **4**: 919–941, doi: 10.5194/gmd-4-919-2011.

Wang X, Parrish D, Kleist D, Whitaker J. 2013. GSI 3DVar-based ensemble-variational hybrid data assimilation for NCEP global forecast system: Single-resolution experiments. *Mon. Weather Rev.* **141**: 4098–4117, doi: 10.1175/MWR-D-12-00141.1.

Wlasak MA, Cullen MJP. 2014. Modelling static 3D spatial background-error covariances – the effect of vertical and horizontal transform order. *Adv. Sci. Res.* **11**: 63–67, doi: 10.5194/asr-11-63-2014.

Wood N, Staniforth A, White AA, Allen T, Diamantakis M, Gross M, Melvin T, Smith C, Vosper S, Zerroukat M, Thuburn J. 2014. An inherently mass-conserving semi-implicit semi-Lagrangian discretization of the deep-atmosphere global non-hydrostatic equations. *Q. J. R. Meteorol. Soc.* **140**: 1505–1520, doi: 10.1002/qj.2235.