Received: 05 March 2018 | Revised: 02 August 2018 | Accepted: 03 September 2018 | Published on: 22 November 2018

DOI: 10.1002/qj.3401

RESEARCH ARTICLE

Quarterly Journal of the Royal Meteorological Society RMetS

A comparison of hybrid variational data assimilation methods for global NWP

Andrew C. Lorenc | Mohamed Jardak

Met Office, Exeter, UK

Correspondence
Andrew C. Lorenc, Met Office, FitzRoy Road, Exeter EX1 3PB, UK.
Email: andrew.lorenc@metoffice.gov.uk

Variational data assimilation methods are reviewed and compared in the Met Office global numerical weather prediction system. This supports hybrid background-error covariances which are a weighted combination of modelled static "climatological" covariances with covariances calculated from a current ensemble of forecasts, in both three- and four-dimensional methods. For the latter, we compare the use of linear and adjoint models (hybrid-4DVar) with the direct use of ensemble forecast trajectories (hybrid-4DEnVar). Earlier studies had shown that hybrid-4DVar outperforms hybrid-4DEnVar, and 4DVar outperforms 3DVar. Improvements in the processing of ensemble covariances and computer enhancements mean we are now able to explore these comparisons for the full range of hybrid weights. We find that, using our operational 44-member ensemble, the static covariance is still beneficial in hybrid-4DVar, so that it significantly outperforms hybrid-4DEnVar. In schemes not using linear and adjoint models, the static covariance is less beneficial. It is shown that the time-propagated static covariance is the main cause of the better performance of 4DVar; when using pure ensemble covariances, 4DVar and 4DEnVar show similar skill. These results are consistent with nonlinear dynamics theory about assimilation in the unstable subspace.

KEYWORDS
4DVar, 4DEnVar, hybrid, NWP

1 | INTRODUCTION

Variational data assimilation (DA) methods have been used for global numerical weather prediction (NWP) since Parrish and Derber (1992). They have contributed significantly to the steady improvement in global NWP, because they can handle efficiently the direct assimilation of large numbers of satellite radiances, and they correct the wide range of scales which can be important. Many centres implemented three-dimensional (3D) variational methods (3DVar)[1] as a stepping stone to 4DVar, which extends this to four dimensions (4D) by iterating integrations of a forecast model and its adjoint within the DA algorithm (Rabier, 2005). Their main competitors were ensemble Kalman filter methods (Kalnay *et al.*, 2007). A review by Lorenc (2003) suggested that the best approach would be a hybrid and outlined a way of doing this. This extended control variable method (also suggested by Buehner, 2005) was implemented in hybrid-4DVar by Clayton *et al.* (2013).

The sequential minimization algorithm of a variational DA method, with each iteration requiring integrations of a forecast model and its adjoint, may make 4DVar less efficient on future massively parallel computers. A variational method (4DEnVar) extending the use of the ensemble to four dimensions, and hence not needing these internal model integrations, was suggested by Liu *et al.* (2008; 2009) and tested in NWP models by Buehner *et al.* (2010a; 2010b; 2015).

Although each method was documented and tested when implemented, there have been few comprehensive comparisons. An exception was Buehner *et al.* (2010a;

[1]This paper follows the nomenclature recommended by Lorenc (2013).

2010b), however they did not evaluate hybrid-4DVar because their linear and adjoint models were obsolescent due to planned changes in forecast model and computers. Lorenc *et al.* (2015) found that, in the Met Office system, hybrid-4DVar performed better than hybrid-4DEnVar. They suggested this was due to the 3D treatment of the static covariance in hybrid-4DEnVar, but they could not test this because of the poor quality of their ensemble. Nevertheless, hybrid-4DEnVar has been selected for operational use in some centres (e.g. Buehner *et al.*, 2015) and is being developed in others (e.g. Desroziers *et al.*, 2014). Hybrid 3DEnVar has been operational at NCEP since 2012, with the switch to hybrid 4DEnVar in January 2015. At the Met Office a new forecast model suitable for exascale supercomputers is expected to be available from 2020, and the development of a new generation of DA software is being planned. In order to define requirements and priorities for this development, we need to evaluate the various options available in the current system.

Lorenc (2017) developed and tested methods for improving the ensemble covariances used in hybrid-variational methods, and they have been further developed and tested for use in the operational hybrid-4DVar (Inverarity *et al.*, 2018). Thanks to these improvements, and a more powerful supercomputer, we are now able to repeat the comparison of Lorenc *et al.* (2015) exploring the full range of hybrid weightings of the static and ensemble covariances. To help understand the differences in performance, we also repeated some of the experiments of Lorenc and Rawlins (2005), comparing 3D and 4D methods. Section 2 defines the hybrid-variational methods compared in this paper and section 3 describes the trialling system and the full set of experiments. Section 4 presents results, first exploring the range of hybrid weightings, then investigates the benefit of time-evolved covariances in 4DVar. Section 5 discusses whether our conclusions using the Met Office NWP system should apply to other systems. Finally section 6 discusses the results and the conclusions about the methods which can be drawn from them.

2 | VARIATIONAL DATA ASSIMILATION METHODS

Bannister (2017) gives a general review of the different variational methods, and our implementation was described in detail by Lorenc *et al.* (2015); here we repeat enough to document the methods as used in this paper and the key differences between them.

2.1 | Four-dimensional best fit

All the methods in this paper perform a 4D best-fit to observations in a time window. We use an underline to extend the standard notation of Ide *et al.* (1997) to four dimensions. The expected error covariance of the background trajectory $\underline{\mathbf{x}}^{\mathrm{b}}$ is $\underline{\mathbf{P}}$. All the methods variationally determine the $\delta\underline{\mathbf{x}}$ which maximizes the posterior Bayesian likelihood by minimizing a penalty function measuring the distances from the background and the observations:

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

where the second term is the observational penalty (J_{o}), measuring the difference, weighted by the inverse observational-error covariance ($\underline{\mathbf{R}}^{-1}$), of the observations in the time window ($\underline{\mathbf{y}}^{\mathrm{o}}$) from their model estimates ($\underline{\mathbf{y}}$). The latter can be calculated using the nonlinear observation operator:

$$\underline{\mathbf{y}}=\underline{H}\left(\underline{\mathbf{x}}^{\mathrm{b}}+\delta\underline{\mathbf{x}}\right). \quad (2)$$

Following the incremental approach (Courtier *et al.*, 1994), most variational systems approximate Equation 2 by

$$\underline{\mathbf{y}}=\underline{H}\left(\underline{\mathbf{x}}^{\mathrm{b}}\right)+\underline{\mathbf{H}}\delta\underline{\mathbf{x}}, \quad (3)$$

where $\underline{\mathbf{H}}$ is tangent-linear to $\underline{H}$. Nonlinearities in $\underline{H}$ are accounted for in an outer loop. The Met Office system was designed from the outset to use highly nonlinear $\underline{H}$, even in 3DVar (Lorenc *et al.*, 2000), so instead we split $\underline{H}$ into linear time- and space-interpolation ($\underline{\mathbf{G}}$) of model-generated fields to the observation times and positions, followed by a nonlinear observation operator (H) operating on these local values:

$$\underline{\mathbf{y}}=H\left(\underline{\mathbf{G}}\,\underline{\mathbf{x}}^{\mathrm{b}}+\underline{\widetilde{\mathbf{G}}}\delta\underline{\mathbf{x}}\right). \quad (4)$$

Using Equation 4 instead of Equation 3 makes Equation 1 non-quadratic, which restricts the choice of minimization algorithms. Nevertheless the minimization can be efficiently performed making a sequence of linear approximations to H[2], moreover this approach reduces the need for outer-loop iterations involving expensive reruns of the forecast model. No outer loop was used for the results reported here (the consequences of this are discussed in section 5 item 3).

We use a different symbol ($\underline{\widetilde{\mathbf{G}}}$) for the increment interpolation because it is implemented differently. We use a lower spatial resolution for $\delta\underline{\mathbf{x}}$, and the time-interpolation differs for each variational method. The background or "first-guess" interpolation ($\underline{\mathbf{G}}\,\underline{\mathbf{x}}^{\mathrm{b}}$) is identical in all the methods in this paper; even in 3D methods we used the so-called first-guess at appropriate time (FGAT). The benefit of FGAT was demonstrated by Lorenc and Rawlins (2005).

For global NWP we cannot evaluate the first term in Equation 1 directly because $\underline{\mathbf{P}}$ is too big. The solution adopted in variational methods is to model $\underline{\mathbf{P}}$ using a sequence of operations we can compute, then use these to transform $\delta\underline{\mathbf{x}}$ so that $(1/2)\delta\underline{\mathbf{x}}^{\mathrm{T}}\underline{\mathbf{P}}^{-1}\delta\underline{\mathbf{x}}$ simplifies. The methods in this paper differ in the transform used to define $\delta\underline{\mathbf{x}}$ and hence the implicit definition of $\underline{\mathbf{P}}$.

[2]In our 'double inner loop' minimization (Rawlins *et al.*, 2007), $\underline{H}$ is linearized about the current best estimate; this linearized $\underline{\mathbf{H}}$ is then used for ten quadratic iterations with exact line searches, before re-linearizing.

It is not immediately obvious how best to add the 4D analysis increment ($\delta\underline{\mathbf{x}}$) to the background ($\underline{\mathbf{x}}^{\mathrm{b}}$) and get the 3D analysis state needed to start the forecast for the next DA cycle. Each method adopts a different approach incorporating techniques for avoiding rapid oscillations in the forecast. Details are in the subsections below.

2.2 | 3DVar

The main feature of 3DVar is its use of a static time-invariant covariance $\mathbf{B}$ to describe the expected error covariance in three dimensions, including balance relationships between variables. We model this using transforms

$$\mathbf{B} = \mathbf{U}\mathbf{U}^{\mathrm{T}}. \tag{5}$$

For these experiments we used the covariance model operational in July 2017 (Wlasak and Cullen, 2014). We construct a 3D analysis increment $\delta\mathbf{x} = \mathbf{U}\mathbf{v}^{\mathrm{c}}$ by transforming a control variable $\mathbf{v}^{\mathrm{c}}$, and transform it to 4D by a persistence forecast model:

$$\delta\underline{\mathbf{x}} = \underline{\mathbf{I}}\delta\mathbf{x}. \tag{6}$$

This gives an implicit 4D prior covariance which is constant in time

$$\underline{\mathbf{P}} = \underline{\mathbf{I}}\mathbf{B}\underline{\mathbf{I}}^{\mathrm{T}} \tag{7}$$

and a transformed penalty function

$$J\left(\mathbf{v}^{\mathrm{c}}\right) = \frac{1}{2}\mathbf{v}^{\mathrm{c}\mathrm{T}}\mathbf{v}^{\mathrm{c}} + \frac{1}{2}\left(\underline{\mathbf{y}} - \underline{\mathbf{y}}^{\mathrm{o}}\right)^{\mathrm{T}}\underline{\mathbf{R}}^{-1}\left(\underline{\mathbf{y}} - \underline{\mathbf{y}}^{\mathrm{o}}\right) \tag{8}$$

The nominal time for the 3D increment $\delta\mathbf{x}$ is the middle of the time window and we assume in Equation 6 that the increment does not vary. To avoid introducing rapid oscillations into the model integrations, instead of adding $\delta\mathbf{x}$ directly to the model background at the nominal analysis time, we add it gradually using an IAU (Incremental Analysis Update) scheme (Bloom *et al.*, 1996). Throughout this paper, we use an IAU scheme to add the increment spread uniformly over a six-hour insertion centred on the increment time, which is the configuration that was used when 3DVar was first made operational at the Met Office in 1999 (Lorenc *et al.*, 2000).

2.3 | 4DVar

To get 4DVar we simply assume that the same transform gives a 3D increment ($\delta\mathbf{x}_0$) valid at the beginning of the time window, and replace persistence by a linear forecast model $\underline{\mathbf{M}}$

$$\delta\underline{\mathbf{x}} = \underline{\mathbf{M}}\delta\mathbf{x}_0\,. \tag{9}$$

This gives an implicit 4D prior covariance

$$\underline{\mathbf{P}} = \underline{\mathbf{M}}\mathbf{B}\underline{\mathbf{M}}^{\mathrm{T}} \tag{10}$$

Since we are running $\underline{\mathbf{M}}$ during the variational minimization, it is straightforward to add an additional J_{c}-term to Equation 8 penalizing rapid variations. Our J_{c}-term is similar to that of Gauthier and Thépaut (2001), but uses a different norm to measure the size of the filter increments.[3] This obviates the need for an IAU to control balance; $\delta\mathbf{x}_0$ can simply be added and the forecast run from the beginning of the window.

To allow a cleaner comparison with 3DVar, we also ran a 4DVar-IAU experiment (discussed in section 4.2) which did not use the J_{c}-term and added the increment valid in the middle of the 6 h window using the IAU.

2.4 | 3DEnVar

All the EnVar and hybrid methods require an ensemble of forecasts, sampling the background pdf. Methods for generating this ensemble, ensuring it is a realistic sample with spread matching the expected background error, are discussed elsewhere (e.g. Bowler *et al.*, 2017b); here we assume that this has been done as well as possible and we concentrate on estimating the best covariance from the necessarily small sample. The methods used were developed and tested in hybrid-4DEnVar by Lorenc (2017). The covariances were improved by using a small number of wavebands (Buehner, 2012), each with a different localization, with scales calculated using the *hybrid-diag* approach (Ménétrier *et al.*, 2015a; 2015b). We augmented the ensemble by lagged perturbations from longer ensemble forecasts (valid at the correct time) from earlier cycles, and by time-shifted perturbations (valid at a different time). The settings in these experiments are documented in section 3; we used four wavebands, two time-lags (i.e. the current ensemble and a longer forecast from the previous cycle) and three time-shifts (i.e. the valid time and perturbations 3 and 6 h later from the current ensemble, earlier from the lagged ensemble). This increased the number of perturbation fields by a factor of 24, however they do not provide new independent samples so are not as effective as a 24-fold increase in ensemble size would be. Lorenc (2017) estimated that the wavebands and the lag-shift approaches are each about as useful as doubling the number of independent ensemble members, together roughly quadrupling it.

To extend the equations of Lorenc *et al.* (2015) to describe this set-up, we can consider each of the derived ensembles (24 in our case), distinguished by a subscript d. Each ensemble of N_{e} members is processed independently, subtracting the ensemble mean, normalizing by $(N_{\mathrm{e}}-1)^{-1/2}$ and applying the waveband filter, to give a matrix of normalized perturbations $\mathbf{X}_d$; each is associated with a localization matrix $\mathbf{L}_d$. The localized covariances are weighted by β_d^2 so that they sum to the improved covariance estimate

$$\mathbf{B}_{\mathrm{ens}} = \sum_d \beta_d^2\,\mathbf{L}_d \circ \mathbf{X}_d\mathbf{X}_d^{\mathrm{T}}. \tag{11}$$

[3] While Gauthier and Thépaut (2001) use the full energy norm, the Met Office scheme uses only the 'elastic' term that depends on the pressure increment, and applies this to both pressure and pressure tendency.

We chose to make the weights β_d^2 equal. Each $\mathbf{L}_d$ (which may have a different localization scale) is modelled using transforms

$$\mathbf{L}_d = \mathbf{U}_d^{\alpha}\,\mathbf{U}_d^{\alpha\mathrm{T}}. \tag{12}$$

The increment $\delta\mathbf{x}$ used in 3DEnVar is a linear combination of ensemble perturbations $\mathbf{x}'_{d,k}$ (i.e. the columns of $\mathbf{X}_d$), localized by weights $\boldsymbol{\alpha}_{d,k}$ which are constrained to be smooth, consistent with Equation 12:

$$\boldsymbol{\alpha}_{d,k} = \mathbf{U}_d^{\alpha}\,\mathbf{v}_{d,k}^{\alpha}, \tag{13}$$

$$\delta\mathbf{x} = \sum_d \beta_d \sum_{k=1}^{N_d} \boldsymbol{\alpha}_{d,k} \circ \mathbf{x}'_{d,k}, \tag{14}$$

using ensembles valid (except for the time-shifts) in the middle of the time window. $\delta\mathbf{x}$ is assumed not to vary over the window; we use Equation 6. This increment is added using the IAU to start the next forecast.

2.5 | Hybrid-3DEnVar

This simply combines 3DVar and 3DEnVar, with overall weights to Equations 5 and 11 such that $\beta_c^2 + \beta_e^2 = 1$

$$\delta\mathbf{x} = \beta_c \mathbf{U}\mathbf{v}^c + \beta_e \sum_d \beta_d \sum_{k=1}^{N_d} \mathbf{U}_e^{\alpha}\mathbf{v}_{d,k}^{\alpha} \circ \mathbf{x}'_{d,k}, \tag{15}$$

$$\underline{\mathbf{P}} = \underline{\mathbf{I}}\left(\beta_c^2\mathbf{B} + \beta_e^2\mathbf{B}_{\mathrm{ens}}\right)\underline{\mathbf{I}}^{\mathrm{T}}. \tag{16}$$

As this is a simple hybrid, the method could equally be called hybrid-3DVar. The experiments reported in section 4.1 used the same software as hybrid-4DEnVar (with the ensemble trajectories containing only one time), so we use the hybrid-3DEnVar name.

2.6 | Hybrid-4DVar

This combines 4DVar with hybrid-3DEnVar, using a covariance $\mathbf{B}_{\mathrm{ens}}$ estimated using ensembles valid at the beginning of the time window

$$\delta\underline{\mathbf{x}} = \underline{\mathbf{M}}\left(\beta_c\mathbf{U}\mathbf{v}^c + \beta_e \sum_d \beta_d \sum_{k=1}^{N_d} \mathbf{U}_e^{\alpha}\mathbf{v}_{d,k}^{\alpha} \circ \mathbf{x}'_{d,k}\right), \tag{17}$$

$$\underline{\mathbf{P}} = \underline{\mathbf{M}}\left(\beta_c^2\mathbf{B} + \beta_e^2\mathbf{B}_{\mathrm{ens}}\right)\underline{\mathbf{M}}^{\mathrm{T}}. \tag{18}$$

The same wavebands and time-lag and shift described in section 2.4 were used. This system was tested in July 2017 as a candidate upgrade for our global hybrid-4DVar (Inverarity *et al.*, 2018). It was implemented operationally (except for the use of time-lagged and shifted ensemble perturbations) in February 2018.

2.7 | Hybrid-4DEnVar

Hybrid-4DEnVar is an extension of hybrid-3DEnVar to use 4D ensemble trajectories. Our implementation persisted the static component as in 3DVar, as well as the $\boldsymbol{\alpha}_{d,k}$ localization fields

$$\delta\underline{\mathbf{x}} = \beta_c\underline{\mathbf{I}}\mathbf{U}\mathbf{v}^c + \beta_e \sum_d \beta_d \sum_{k=1}^{N_d} \left(\underline{\mathbf{I}}\mathbf{U}_e^{\alpha}\mathbf{v}_{d,k}^{\alpha}\right) \circ \underline{\mathbf{x}}'_{d,k}, \tag{19}$$

$$\underline{\mathbf{B}}_{\mathrm{ens}} = \sum_d \beta_d^2 \left(\underline{\mathbf{I}}\mathbf{L}_d\underline{\mathbf{I}}^{\mathrm{T}}\right) \circ \left(\underline{\mathbf{X}}_d\underline{\mathbf{X}}_d^{\mathrm{T}}\right), \tag{20}$$

$$\underline{\mathbf{P}} = \beta_c^2\underline{\mathbf{I}}\mathbf{B}\underline{\mathbf{I}}^{\mathrm{T}} + \beta_e^2\underline{\mathbf{B}}_{\mathrm{ens}}. \tag{21}$$

Since Equation 19 generates a 4D increment $\delta\underline{\mathbf{x}}$, we generalize the IAU to a 4DIAU (Lorenc *et al.*, 2015), each timestep adding a fraction of the 3D component of $\delta\underline{\mathbf{x}}$ which is closest to the current time of the forecast model. In these experiments the ensemble trajectories were saved every 3 h, so the 4D IAU added the fields at the beginning, middle and end of the assimilation window with constant IAU weights for 1.5, 3 and 1.5 h respectively.

2.8 | Hybrid-3DVar-$\mathbf{M}_3\mathbf{B}\mathbf{M}_3^{\mathrm{T}}$

This method uses the hybrid-4DVar software to run a 3DVar experiment with the implicit background covariance evolved. The method, deliberately designed to be between 3DVar and 4DVar, was conceived for a diagnostic experiment by Lorenc and Rawlins (2005), who called it synoptic 4DVar. It is not suggested as an operational method since hybrid-4DVar should be better. The linear model $\mathbf{M}_3$ is integrated from a control increment at 0 h to 3 h (times relative to the start of the window) to create a single analysis increment state at the middle of the window (instead of the 4D trajectory created by $\underline{\mathbf{M}}$); as in 3DVar (FGAT) all observations in the time window are fitted using this single increment – in effect the trajectory needed for Equation 2 is created by a persistence forecast $\underline{\mathbf{I}}$:

$$\delta\underline{\mathbf{x}} = \underline{\mathbf{I}}\mathbf{M}_3\left(\beta_c\mathbf{U}\mathbf{v}^c + \beta_e \sum_d \beta_d \sum_{k=1}^{N_d} \mathbf{U}_e^{\alpha}\mathbf{v}_{d,k}^{\alpha} \circ \mathbf{x}'_{d,k}\right), \tag{22}$$

$$\underline{\mathbf{P}} = \underline{\mathbf{I}}\mathbf{M}_3\left(\beta_c^2\mathbf{B} + \beta_e^2\mathbf{B}_{\mathrm{ens}}\right)\mathbf{M}_3^{\mathrm{T}}\underline{\mathbf{I}}^{\mathrm{T}}. \tag{23}$$

In order to fully match 3DVar, the J_c term is not used; the analysis increment is added using the IAU.

Note that the $\mathbf{B}_{\mathrm{ens}}$ used in this scheme is valid, like that in hybrid-4DVar, at the beginning of the window. So it needs time evolution in Equation 23 to give an increment valid at the middle. In contrast $\mathbf{B}_{\mathrm{ens}}$ in the hybrid-3DEnVar method (Equation 16) is evaluated directly from the ensemble forecasts valid in the middle of the window.

3 | TRIALS

All the experiments in Table 1 were run using N320 grids (approximately 40 km in the horizontal at UK latitudes), 70 levels (extending to 80 km vertically) for the model forecasts and N216 (approximately 60 km in the horizontal) analysis increments.

The ensemble system was a lower-resolution version of that operational in July 2017, with 44 members generated

TABLE 1 List of trials performed

Trial and symbol on figures		Software configuration	Initialization method	β_c^2	β_e^2	% change in NWP Index	Initial $J_o \times 10^{-4}$	Final $J_o \times 10^{-4}$
4DVar	□	4DVar	J_c	1	0	0.00	970	690
Hybrid-4DVar	□	4DVar	J_c	0.7	0.3	0.98	962	686
Hybrid-4DVar	□	4DVar	J_c	0.5	0.5	1.07	963	691
Hybrid-4DVar	□	4DVar	J_c	0.3	0.7	0.98	968	700
4DVar-$\mathbf{B}_{ens}$	□	4DVar	J_c	0	1	−1.26	1014	751
$3DVar_{EnVar}$	◊	EnVar	IAU	1	0	−5.92	1097	686
Hybrid-4DEnVar	◊	EnVar	4D IAU	0.7	0.3	-1.96	1040	664
Hybrid-4DEnVar	◊	EnVar	4D IAU	0.5	0.5	−1.52	1031	668
Hybrid-4DEnVar	◊	EnVar	4D IAU	0.3	0.7	−1.20	1025	676
4DEnVar	◊	EnVar	4D IAU	0	1	−1.77	1063	726
$3DVar_{EnVar}$	★					*see above*		
Hybrid-3DEnVar	★	EnVar	IAU	0.7	0.3	−2.46	1056	679
Hybrid-3DEnVar	★	EnVar	IAU	0.5	0.5	−2.08	1049	682
Hybrid-3DEnVar	★	EnVar	IAU	0.3	0.7	−2.03	1047	686
3DEnVar	★	EnVar	IAU	0	1	−2.62	1106	724
3DVar-$\mathbf{M}_3\mathbf{B}\mathbf{M}_3^{\mathrm{T}}$	△	4DVar	IAU	1	0	−1.64	1046	714
3DVar-$\mathbf{M}_3\mathbf{B}\mathbf{M}_3^{\mathrm{T}}$	△	4DVar	IAU	0.7	0.3	−0.68	1027	709
3DVar-$\mathbf{M}_3\mathbf{B}\mathbf{M}_3^{\mathrm{T}}$	△	4DVar	IAU	0.5	0.5	−0.38	1028	714
3DVar-$\mathbf{M}_3\mathbf{B}\mathbf{M}_3^{\mathrm{T}}$	△	4DVar	IAU	0.3	0.7	−0.68	1032	723
3DVar-$\mathbf{M}_3\mathbf{B}\mathbf{M}_3^{\mathrm{T}}$	△	4DVar	IAU	0	1	−3.34	1100	769
4DVar-noJ_c	×	4DVar	none	1	0	−0.03	979	690
4DVar-IAU	×	4DVar	IAU	1	0	−0.46	977	689
4DVar-obst_3	×	4DVar	J_c	1	0	−0.81	1019	719
3DVar-$\mathbf{M}_6\mathbf{B}\mathbf{M}_6^{\mathrm{T}}$	×	4DVar	IAU	1	0	−1.18	1041	724
3DVar-$\mathbf{M}_1\mathbf{B}\mathbf{M}_1^{\mathrm{T}}$	×	4DVar	IAU	1	0	−4.21	1078	696
3DVar-$\mathbf{M}_f\mathbf{B}\mathbf{M}_f^{\mathrm{T}}$	×	4DVar	IAU	1	0	−6.91	1113	755
3DVar	×	4DVar	IAU	1	0	−5.84	1096	679

by a recentred Local Ensemble Transform Kalman Filter (LETKF; Bowler *et al.*, 2008; Flowerdew and Bowler, 2011). Operationally there is a two-way coupling between the DA system, which uses ensemble forecasts in its hybrid (EnVar) component, and the ensemble system, which recentres the ensemble about the DA analysis. For simplicity and cost reasons, these experiments were only one-way coupled; they all used the same ensemble forecasts, recentred about the operational DA. All the experiments also used the same observational bias-correction coefficients, calculated with the operational variational bias-correction system. Removing the recentring might be expected to have a strong effect, but is not possible in the ensemble system we used. There is no evidence that changing from a frozen (operational) recentring to full two-way coupled recentring is worthwhile for trials like these. Because of possible differences in systematic errors, the variational bias correction would also need to be coupled, requiring longer trial periods with a more expensive fully coupled system.

The ensemble forecasts were reconfigured to the analysis resolution before use. The perturbations from the ensemble mean were transformed and filtered into four wavebands, as in Lorenc (2017). Gaussian-shaped horizontal localization functions were used, with scales 6241, 919, 389, 256 km for the bands. These localization scales, and the vertical localization matrix, were estimated using the *hybrid-diag* approach (Ménétrier *et al.*, 2015a; 2015b). As explained in section 2.4, a six-fold increase in ensemble size was obtained using lags=[0,1] cycles, validity-time-shifts=[0,3,6] h.

Our software can be run in different configurations, which have some differences in detail:

4DVar Because of the cost of running the linear and adjoint models at high resolution, we first calculate preconditioning in 40 low-resolution (N108) iterations, then limit the minimization to 33 high-resolution (N216) iterations. The variable transforms $\mathbf{U}$ and the models are linearized about the background trajectory $\underline{\mathbf{x}}^{\mathrm{b}}$, interpolated to the analysis grid.

EnVar The cost is less, so we can afford 60 iterations at N216, getting quite close to full convergence. The variable transforms $\mathbf{U}$ are linearized about the ensemble mean trajectory.

All the trials were run for seven weeks from 0600 UTC on 15 May 2015 to 0000 UTC on 3 July 2015, with forecasts to 5 days every 12 h. As we have run many trials, we focus

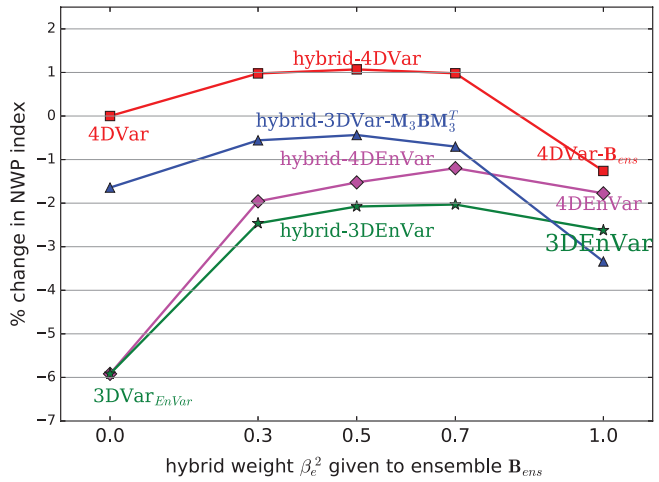

FIGURE 1 Change in NWP Index, relative to 4DVar, for the hybrid-4DVar (red □), hybrid-4DEnVar (magenta ◇), hybrid-3DEnVar (green ★) and hybrid-3DVar-$\mathbf{M}_3\mathbf{B}\mathbf{M}_3^{\mathrm{T}}$ (blue △) methods documented in section 2. The change in Index (defined such that positive values indicate a better forecast) for each method is plotted as a function of the hybrid weight β_e^2 given to the ensemble covariance $\mathbf{B}_{\mathrm{ens}}$ [Colour figure can be viewed at wileyonlinelibrary.com]

discussion first on a single summary score (shown in column 7 of Table 1) – the expected percentage change in "NWP Index" (Bowler *et al.*, 2017a, appendix) compared to a control experiment; we have chosen to use 4DVar as control. The Index is based on a weighted combination of the root mean square (RMS) errors verified against observations, for a selection (shown in Figure 2) of forecast lengths and fields. Each is converted to a skill score using the corresponding persistence forecast, then weighted according to its judged importance to our forecast users. Over the past five years, the Index has increased by a little over 1% per year, so an increase of 1% from a trial is a substantial improvement. Our results are averaged over 42 days, the trial period omitting the first week.

Thanks to a recent computer enhancement, we were able to run our main trials exploring the full range of hybrid weights in experiments with $\beta_e^2 = [0.0, 0.3, 0.5, 0.7, 1.0]$. Similar studies, showing how a hybrid performs better than pure static or ensemble covariance, have been published for simplified models (Etherton and Bishop, 2004), but not for full NWP systems. These results, from the top four groups in Table 1, are presented and discussed in section 4.1. They highlighted the large difference in performance between 4DVar and 3DVar (both with $\beta_e^2 = 0.0$), so additional trials (in the bottom group) studying this are presented in section 4.2.

4 | RESULTS

4.1 | Trials exploring hybrid weights

Figure 1 shows results for the three main methods compared by Lorenc *et al.* (2015) (hybrid-4DVar, hybrid-4DEnVar and hybrid-3DEnVar), for the full range of hybrid weights from pure static covariances on the left to pure ensemble covariances on the right. (The hybrid-3DVar-$\mathbf{M}_3\mathbf{B}\mathbf{M}_3^{\mathrm{T}}$ results will be discussed later.) Because our implementation of hybrid-4DEnVar uses a fixed static covariance, it tends on the left to the same result as hybrid-3DEnVar; the limit is labelled $3\mathrm{DVar}_{\mathrm{EnVar}}$ since with $\beta_e^2 = 0.0$ it should not be called a hybrid or ensemble method, but it used the EnVar software. At the other extreme, the hybrid prefix is again dropped; the experiments are called 4DEnVar and 3DEnVar. Following the nomenclature of Lorenc (2013), the extreme case of hybrid-4DVar, with $\beta_e^2 = 1.0$, might be called En-4DVar, but this name is already used elsewhere for ensembles of 4DVar assimilations. So this paper follows Buehner *et al.* (2010b) and calls it 4DVar-$\mathbf{B}_{\mathrm{ens}}$.

Figure 1 confirms and expands the conclusion of Lorenc *et al.* (2015): it shows that the best hybrid-4DVar has $\beta_e^2 = 0.5$ and significantly outperforms the best hybrid-4DEnVar with $\beta_e^2 = 0.7$. Figure 2 gives more detail on the regional distribution of this comparison. Figure 2a,b show that hybrid-4DVar improves on 4DVar in all regions and timescales, except day 5 in the Northern Hemisphere Extratropics. (The day 5 scores are the most noisy in trials; they may not be reliable in our 42 day average.) All the fields and regions give some contribution to the overall improvement of 1.1% in the NWP Index. Figure 2c,d show the improvement of hybrid-4DVar over hybrid-4DEnVar. Again, nearly all fields and regions contribute to a difference in NWP Index of 2.3%; the biggest contributions come from tropical winds, and also about 0.3% from day 5 (this might be affected by sampling noise).

The comparison of 4DEnVar with 3DEnVar in Figure 1 indicates that allowing for the time of the innovations within the window (in the time-interpolation $\underline{\widetilde{\mathbf{G}}}$ in Equation 4) improves the NWP Index by nearly 1%. At the other extreme, we see that 4DVar gives a very large improvement over 3DVar of nearly 6%. It seems that the 4DVar algorithm, involving a linear model and its adjoint, improves the static $\mathbf{B}$ considerably, so that 4DVar using this outperforms 4DVar-$\mathbf{B}_{\mathrm{ens}}$. In contrast, 4DEnVar makes better use of $\underline{\mathbf{B}}_{\mathrm{ens}}$; the hybrid adding some static $\mathbf{B}$ performs only slightly better. 3DVar using $\mathbf{B}$ without any linear models is much worse than 3DEnVar using $\mathbf{B}_{\mathrm{ens}}$.

To investigate this effect of the linear and adjoint model further, we ran a further set of experiments inspired by a similar experiment in Lorenc and Rawlins (2005), and the pioneering study of Thépaut *et al.* (1996). The goal is to attribute the improvement from 4DVar to two differences: the treatment of innovations at their correct time and the use of time-evolved covariances. The extra experiments, labelled hybrid-3DVar-$\mathbf{M}_3\mathbf{B}\mathbf{M}_3^{\mathrm{T}}$ in Figure 1, used time-evolved covariances in a 3DVar algorithm, as documented in section 2.8. The difference between the hybrid-3DVar-$\mathbf{M}_3\mathbf{B}\mathbf{M}_3^{\mathrm{T}}$ and hybrid-4DVar results shows that allowing for the time of the innovations within the window (in the time-interpolation $\underline{\widetilde{\mathbf{G}}}$ in Equation 4) improves the NWP Index by over 1% – slightly more than the difference between 4DEnVar and

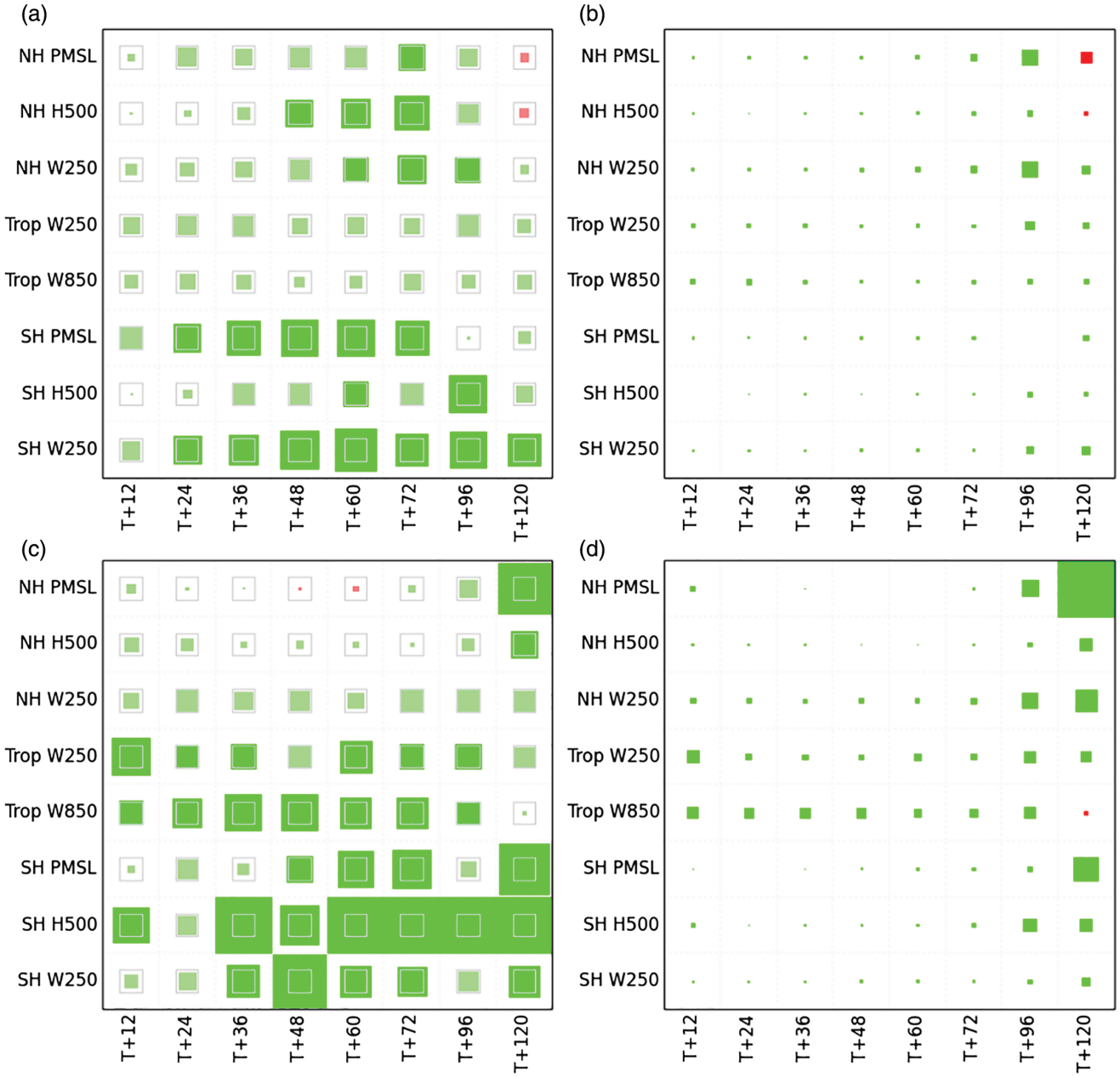

FIGURE 2 Scorecards showing regional improvements due to the best hybrid-4DVar with $\beta_e^2 = 0.5$. (a, b) Contributions to the 1.1% average compared to 4DVar. (c, d) Contributions to the 2.3% average compared to best hybrid-4DEnVar with $\beta_e^2 = 0.7$. Each panel shows boxes with sides proportional to the change (green positive, red negative) with different fields (sea-level pressure PMSL, geopotential height H or wind W at the hPa level shown) and regions (Southern Extratropics SH, Tropics Trop, Northern Extratropics NH) in the y-direction and forecast lengths (hours) in the x-direction. In (a, c) the percentage change in RMS error is shown by a grey line for 2% and by a full box for 5% or more. (b, d) show the corresponding contributions to the change in NWP Index, with a full box denoting 0.1 or more [Colour figure can be viewed at wileyonlinelibrary.com]

3DEnVar. Hybrid-3DVar-$\mathbf{M}_3\mathbf{B}\mathbf{M}_3^{\mathrm{T}}$ copies the other 3D methods by adding increments valid at the middle of the time window using the IAU. We shall see from the 4DVar-IAU experiment discussed in section 4.2 that about 0.5% of the difference just discussed is in fact due to the different initialization methods. So these results are consistent with the 4DEnVar–3DEnVar difference.

As discussed in section 2.8, hybrid-3DVar-$\mathbf{M}_3\mathbf{B}\mathbf{M}_3^{\mathrm{T}}$ uses $\mathbf{M}_3\mathbf{B}_{\mathrm{ens}}\mathbf{M}_3^{\mathrm{T}}$ with $\mathbf{B}_{\mathrm{ens}}$ evaluated at the beginning of the window, whereas hybrid-3DEnVar uses $\mathbf{B}_{\mathrm{ens}}$ evaluated directly at the middle of the window. We see in Figure 1 that with $\beta_e^2 = 1.0$ the latter is better – not surprisingly evolution using the full model give slightly better results than using the approximate linear model. In contrast, comparing $\beta_e^2 = 1.0$ values of hybrid-4DVar and hybrid-4DEnVar, the former gives slightly better results. This may be due to other differences between these experiments in our system, as discussed in section 5, items 6 and 7.

At the other $\beta_e^2 = 0.0$ extreme, the use of a time-evolved $\mathbf{M}_3\mathbf{B}\mathbf{M}_3^{\mathrm{T}}$ gives a large part of the benefit of 4DVar over 3DVar (in agreement with the result of Lorenc and Rawlins, 2005), and so performs better than 3DEnVar.

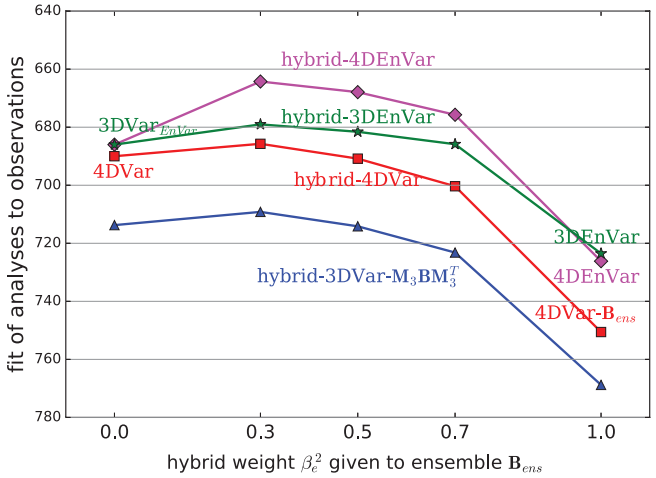

FIGURE 3 As Figure 1, but showing the six-week average fit of the analyses to the observations. This is measured by the final value of J_o, at the end of each variational minimization. The y-axis is scaled by 10^{-4} and reversed so that upwards denotes a better fit [Colour figure can be viewed at wileyonlinelibrary.com]

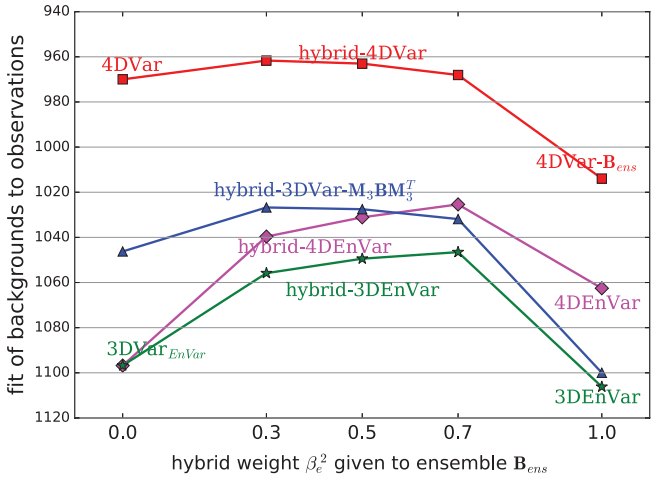

FIGURE 4 As Figure 3, but showing the six-week average fit of the backgrounds to the observations [Colour figure can be viewed at wileyonlinelibrary.com]

From its method of construction, we would expect the static **B** to provide a fairly complete sampling of all likely error patterns, whereas, despite the efforts of Lorenc (2017) to improve it, $\mathbf{B}_{\text{ens}}$ is expected to remain undersampled; many of the actual error patterns seen in the innovations will not be as well fitted using the available ensemble perturbations. This is partly confirmed by a worse fit of analyses to observations for methods with $\beta_e^2 = 1.0$, as seen in Figure 3. On the other hand, the static **B** does not give enough weight to the growing modes which are believed to form a large part of actual errors. These are often tilted along baroclinic surfaces (Thépaut *et al.*, 1996) – our method of modelling **B** favours isotropic non-sloping structures. This latter problem with **B** dominates in hybrid-4DEnVar; best results are obtained with quite a large weight ($\beta_e^2 = 0.7$ in Figure 1) to $\mathbf{B}_{\text{ens}}$. But the use of a time-evolved $\underline{\mathbf{M}}\mathbf{B}\underline{\mathbf{M}}^{\mathrm{T}}$ in hybrid-4DVar largely corrects the shortcomings of **B**; then the better sampling of the static **B**

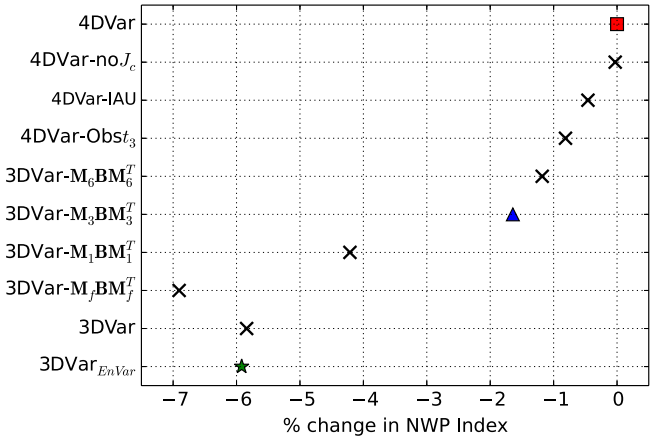

FIGURE 5 Change in NWP Index from extra experiments with $\beta_e^2 = 0$ (×). The $\beta_e^2 = 0$ values from the main hybrid experiments are also plotted using the same symbols as in Figure 1. These results are discussed in section 4.2 [Colour figure can be viewed at wileyonlinelibrary.com]

means best results are obtained when it has more weight and $\mathbf{B}_{\text{ens}}$ has less ($\beta_e^2 = 0.5$ in Figure 1).

Having a good enough fit of the analyses to the observations is a necessary but not sufficient measure of a good assimilation method, so there is little correlation between the scores plotted in Figure 3 and the NWP Index shown in Figure 1. In contrast, the fit of the background (i.e. the short forecast from the previous analysis) to the observations (Figure 4) is very similar to Figure 1.

4.2 | Additional trials investigating 4DVar and 3DVar differences

Figure 5 repeats the $\beta_e^2 = 0$ values for the four curves in Figure 1 alongside other experiments designed to investigate further the 3DVar-$\mathbf{M}_3\mathbf{B}\mathbf{M}_3^{\mathrm{T}}$ results. The experiments are explained and discussed in turn below:

4DVar This is repeated from Figure 1; it is the control against which the other experiments were compared.

4DVar-no$\boldsymbol{J_c}$ This experiment simply omitted the J_c term and added the increment ($\delta\mathbf{x}_0$) at the start of the window with no IAU. It can be compared with 4DVar, which used the J_c term, and 4DVar-IAU. It seems that the J_c term is not giving much benefit in the current system.

4DVar-IAU This is like 4DVar-noJ_c, but adds $\delta\mathbf{x} = \mathbf{M}_3\delta\mathbf{x}_0$ by the IAU. Using the IAU to add the 4DVar increment degrades results by nearly 0.5%. Similar experiments (not shown) with hybrid covariances showed a similar degradation. Lei and Whitaker (2016) also found a degradation when applying an IAU with the NCEP model; the only IAU configuration that improved background fits to observations and longer forecasts was the 4DIAU. Note that all our 3D experiments used the

IAU and probably suffered a similar detriment.

4DVar-Obst_3 This is like 4DVar, but with all the innovations treated as if they were in the middle of the window; i.e. in Equation 4 the time interpolation $\underline{\mathbf{G}}$ is still using the correct time, whereas $\underline{\widetilde{\mathbf{G}}}$ uses the increment from the middle of the time window. This experiment differs from the 3DVar-$\mathbf{M}_3\mathbf{B}\mathbf{M}_3^{\mathrm{T}}$ experiment below it because it uses the J_c term for initialization and adds $\delta\mathbf{x}_0$ instead of adding $\delta\mathbf{x} = \mathbf{M}_3\delta\mathbf{x}_0$ by the IAU. These differences are the same as those between 4DVar and 4DVar-IAU, and have a similar effect on scores.

3DVar-$\mathbf{M_3}\mathbf{B}\mathbf{M}_3^{\mathbf{T}}$ This is repeated from Figure 1; it is a combination of the changes tested in 4DVar-Obst_3 and 4DVar-IAU. This result shows that the effect of the above changes are approximately additive.

3DVar-$\mathbf{M_6}\mathbf{B}\mathbf{M}_6^{\mathbf{T}}$ This is like 3DVar-$\mathbf{M}_3\mathbf{B}\mathbf{M}_3^{\mathrm{T}}$, but using a 6 h evolution of covariances (from 0 to 6 h in the window). The similarity of this with 3DVar-$\mathbf{M}_3\mathbf{B}\mathbf{M}_3^{\mathrm{T}}$ shows that much of the benefit from evolution of covariances in achieved within 3 h.

3DVar-$\mathbf{M_1}\mathbf{B}\mathbf{M}_1^{\mathbf{T}}$ This is like 3DVar-$\mathbf{M}_3\mathbf{B}\mathbf{M}_3^{\mathrm{T}}$ but using a 1 h evolution of covariances (from 0 to 1 h in the window). This comparison shows that 1 h is not long enough to get the full benefit from evolved covariances. This rules out the rapid adjustments which can occur in the linear model's non-hydrostatic dynamics as the main factor in the benefit.

3DVar-$\mathbf{M}_f\mathbf{B}\mathbf{M}_f^{\mathbf{T}}$ This is like 3DVar-$\mathbf{M}_3\mathbf{B}\mathbf{M}_3^{\mathrm{T}}$ but using a 3 h evolution of covariances with a linearization state taken from the start of the experiment (before the verification period). This result demonstrates that evolution of the covariances by the actual flow "of the day" is necessary to get the benefit; evolution by the wrong flow actually slightly degrades the 3DVar results. It also suggests that the benefit of evolution does not come from the selective damping, independent of linearization state, of certain types of motion by the model $\mathbf{M}$.

3DVar This experiment used the method documented in section 2.2, with no evolution of covariances.

3DVar$_{\mathbf{EnVar}}$ This experiment, repeated from Figure 1, used the hybrid-3DEnVar software with $\beta_e^2 = 0$. The 3DVar result above was obtained using the 4DVar software with no linear forecast. Their comparison here is included mainly as a check that our software is behaving correctly, but there is a subtle difference between the runs: the variable transforms $\mathbf{U}$ are linearized about the background $\mathbf{x}^{\mathrm{b}}$ in 3DVar but about the ensemble mean in 3DVar$_{\mathrm{EnVar}}$. This similarity of results confirms as expected that this small change of linearization state is insignificant.

5 | WILL OUR RESULTS APPLY TO OTHER NWP SYSTEMS?

We need to consider how applicable our findings are to other systems, with different detailed designs of 4DVar and 4DEnVar. So below we discuss the main distinguishing features of the Met Office system we used, and how they might affect the conclusions.

1. The covariance model for the static $\mathbf{B}$ allows very little spatial variability in the covariances – less for instance than the wavelet-based approach of Fisher (2003) used in several centres. We do not allow the flow-dependent background-error variances of Raynaud *et al.* (2011). It is possible that the improvement to $\mathbf{B}$ from evolution comes partly by putting in some of the missing spatial variability. It would be interesting to compare 3DVar and 4DVar in a modern NWP system with these improved covariance models; unfortunately no such comparison has been published.
2. All our experiments used a 44-member ensemble, with effective size perhaps doubled using time-lagging and shifting. The operational upgrade in 2013 from 22 to 44 members gave about 0.6% improvement in NWP Index, keeping the same hybrid-4DVar settings. Limited testing has indicated that both this larger ensemble, and the scale-dependent waveband localization, are necessary for the reasonable results in this paper with 100% weight to the ensemble ($\beta_c^2 = 0, \beta_e^2 = 1$). Our results have shown that it is the static covariance component which has an advantage in 4DVar; presumably as the ensemble improves, the optimal β_c^2 will decrease and so will the advantage of 4DVar. Increases in our ensemble size are being tested, so we plan future work to investigate this effect.
3. Our 4DVar does not use an outer loop, instead allowing for nonlinear H in its inner loop. Most centres use an outer loop and linear $\mathbf{H}$. If anything, one might expect this feature to harm our 4DVar results, so it probably did not affect our conclusion that hybrid-4DVar was better than hybrid-4DEnVar. We note that an outer loop is possible in our 4DVar system; the Met Office intends to test it again. Adding an outer loop to 4DEnVar would be more difficult, since we would have to replace the 4D IAU by

another method for initializing the increment at the beginning of the time window. So our use of nonlinear H with no outer loop probably helps 4DEnVar.

4. The linear model (and its corresponding adjoint) used in incremental 4DVar was designed from the outset to give a cheap but reasonably accurate forecast for a range of finite perturbations with amplitude the order of typical analysis increments. It has simplified parametrizations (Stiller and Ballard, 2009; Stiller, 2009; Payne, 2009) which were not directly based on those in the full model. This is in contrast to the approach used in most other 4DVar systems; it may have a bearing on the lack of benefit from an outer loop discussed in 3 above. As there, any weakness in our hybrid-4DVar system would hinder it in our comparison with hybrid-4DEnVar.
5. Our 4DEnVar applies spatial localization to transformed perturbation variables: streamfunction, velocity potential, unbalanced pressure, moisture. This was intended to avoid generation of imbalance (Kepert, 2009), but Clayton *et al.* (2013) found that localization of streamfunction and velocity potential, which have some large-scale features, can generate spurious short-scale noise. To avoid this, they introduced a high-pass filter for the perturbations, which reduced the noise but significantly altered the power spectrum of $\mathbf{B}_{\mathrm{ens}}$. Since the implementation of wavebands, with scale-dependent horizontal localization, these problems with large scales have gone away and we no longer need the high-pass filter. The method for diagnosing the localization scale for each waveband takes account of the covariance scale of the variables in the waveband being localized, and so reduces any differences due to our localization of different variables.
6. Our J_{c} term was thoroughly tuned and tested before implementation in 4DVar in 2005, giving about 0.6% improvement in the NWP Index. Our IAU settings, giving constant equal weight over a 6 h window, were tested and found to be beneficial before implementation with 3DVar in 1999. Our 4D IAU method, chosen mainly for simplicity to be similar to the IAU, was shown to be beneficial by Lorenc *et al.* (2015). Since these various methods were tuned, there have been very significant improvements in NWP; the backgrounds are more accurate and the analysis increments correspondingly smaller. Studies to derive estimates of $\mathbf{B}$ show that errors also have smaller horizontal scales and are less balanced. Our results from 4DVar-noJ_{c} were very similar to those from 4DVar with our old J_{c} settings: this probably indicates that there is now much less need for initialization, although it might be that a retuned J_{c} would be more beneficial. Probably for similar reasons, our results from 4DVar-IAU indicate that the IAU is harming results (at least from systems where the analysis increments have come from our linearized model) by about 0.5%. It is possible that the results for experiments using our hybrid-4DEnVar, hybrid-3DEnVar and hybrid-3DVar-$\mathbf{M}_3\mathbf{B}\mathbf{M}_3^{\mathrm{T}}$ schemes could be improved by this amount by tuning the IAU weights. This would not change the main conclusions.
7. Like most 4DEnVar systems, we used a persistence forecast for the localization matrix. Lorenc *et al.* (2015) showed this could limit the impact of observations in jet streams, if their impact is advected beyond the localization scale, which prompted Desroziers *et al.* (2016) to develop the remedy of advecting the localization. The experiments here used scale-dependent localization, so that large scales in the ensemble covariances use a very large localization scale (6241 km), well beyond the range of any advection. So only smaller scales in 4DEnVar are now affected. The difference in performance (with $\beta_{\mathrm{c}}^2 = 0$) of 4DVar-$\mathbf{B}_{\mathrm{ens}}$, which advects the localized $\mathbf{B}_{\mathrm{ens}}$, and 4DEnVar, which applies fixed spatial localization to a 4D $\underline{\mathbf{B}}_{\mathrm{ens}}$, is small and may be due to the initialization differences, suggesting that the advection of localization is not a large problem for our 6 h window.
8. Like most hybrid-4DEnVar systems, the static covariance component $\mathbf{B}$ is held constant; in the limit $\beta_{\mathrm{e}}^2 = 0$ it behaves like 3DVar. In principle we could advect $\mathbf{B}$ as in 4DVar to give $\underline{\mathbf{M}}\mathbf{B}\underline{\mathbf{M}}^{\mathrm{T}}$, but this require the linear and adjoint models and so defeats a major reason for prefering 4DEnVar. Desroziers *et al.* (2016) suggest using a simpler advective model. As discussed in section 6, perhaps the amplification of unstable modes is important in the improvement of 4DVar over 3DVar; it is unlikely that simple advection would capture this. We conclude that this aspect of our scheme is similar to those used elsewhere.
9. To our knowledge, our system remains the only operational hybrid-4DVar for global NWP, where the ensemble covariances are used directly. ECMWF use their ensemble to train a revised $\mathbf{B}$ using an isotropic covariance model which may not preserve the weight given to the unstable sub-space, as discussed in section 6. Buehner *et al.* (2015) implemented hybrid-4DEnVar in the Canadian NWP system on the basis of comparison with an older (non-hybrid) 4DVar; hybrid-3DEnVar was implemented in the USA system on the basis of comparison with 3DVar (Wang *et al.*, 2013); it was later upgraded to hybrid-4DEnVar.

Taking the above into account, we expect our general conclusions about the different variational methods to be applicable to other global systems.

6 | DISCUSSION AND CONCLUSIONS

Our results have shown that hybrid-4DVar has a significant advantage over hybrid-4DEnVar when the weight given to the static covariance $\mathbf{B}$ is large. The difference is about equal to two years of NWP improvements; note however that potential changes to the hybrid DA tested in the last two years (Bowler *et al.*, 2017a; Lorenc, 2017) improved both methods, so did not significantly reduce the difference.

The result seems to be due to the combination of two effects: the ensemble is not big enough to properly sample all the errors which might occur, and the static **B** component of the hybrid, which is designed to counteract this, performs much better when implicitly evolved in 4DVar. Our model for **B**, like most others, is isotropic – it cannot give preference to modes sloping along baroclinic surfaces. Evolution by a linear model improves this. Both effects are consistent with the concepts of assimilation in the unstable subspace (Trevisan *et al.*, 2010). Bocquet and Carrassi (2017) show that, for an ideal ensemble Kalman filter using a perfect model, the error covariance matrix is asymptotically supported by the unstable–neutral subspace only, and that an ensemble size at least as big as the dimension of the unstable–neutral subspace is needed. Their studies only provide a limiting case; because of the butterfly effect, NWP models cannot be perfect and in practice ensemble covariances are localized to increase the effective ensemble size. Nevertheless they give some understanding as to why our 44-member ensemble covariance $\mathbf{B}_{\text{ens}}$, even after improvement using waveband localization and time-lagging and shifting, is not good enough to fully replace the static **B** in 4DVar, and why **B** is improved by evolution with a linear model which will amplify unstable directions.

Our results give no indication of how big an ensemble is needed to remove the benefit from the static **B** in hybrid-4DVar or hybrid-4DEnVar. Buehner *et al.* (2015), with a 256-member ensemble, said that going beyond equal weights ($\beta_c^2 = \beta_{rme}^2 = 0.5$) did not give any improvement in the Environment Canada global hybrid-4DEnVar system, suggesting that some aspect of their ensemble covariances requires improvement before they can be given higher weight. The NCEP global hybrid-4DEnVar system uses an 80-member ensemble and $\beta_c^2 = 0.125$, $\beta_e^2 = 0.875$ (Kleist, personal communication, 2017). In EnKF systems (without any static covariances) Houtekamer *et al.* (2014) have shown benefit from ensembles with up to 384 members, and research systems are testing much larger ensembles (Miyoshi *et al.*, 2014).

Our results showed a large (6%) improvement of 4DVar over 3DVar; this is a major factor in the advantage of hybrid-4DVar over hybrid-4DEnVar. When 4DVar was first implemented in 2004 the improvement over 3DVar was only 2.6%, a similar impact to 4DVar implementations at other centres (Rawlins *et al.*, 2007). At that time we used a basic 4DVar system with all other settings left as they had been developed for 3DVar. Since then there have been substantial developments, such as the linear model parametrizations and the observation usage, which were shown to improve the 4DVar system. The IAU initialization used by 3DVar now seems to degrade results by perhaps 0.5%. Only an approximate comparison is possible because the NWP Index was defined somewhat differently in 2004. So our current 6% difference cannot be said to be inconsistent with earlier results.

Another remarkable feature of the current results is that only 3 h of evolution of static **B** covariances produces most of the benefit – the 3DVar-$\mathbf{M}_6\mathbf{B}\mathbf{M}_6^{\mathrm{T}}$ experiment with 6 h evolution was only a little better (Figure 5). This contrasts with the toy model results of Fisher *et al.* (2005) which suggested that windows of up to 4 days might be beneficial. A period of 3 h is not long enough to asymptote to the unstable–neutral subspace, so we might speculate that while **B** is deficient in this space, other error mechanisms are also important. In our 4DVar, $\underline{\mathbf{M}}$ forecasts from 0 to 6 h, so we chose the average length for the 3-hour evolution $\mathbf{M}_3$. ECMWF use a 12 h window (Bonavita *et al.*, 2016), but for operational short-period NWP there is an incentive to try shorter windows (Payne, 2017). The poor 3DVar-$\mathbf{M}_1\mathbf{B}\mathbf{M}_1^{\mathrm{T}}$ results suggest that windows shorter than about 6 h lose some of the advantages of 4DVar.

In contrast to 4DVar, we found a much smaller benefit from adding a hybrid static **B** to 3DEnVar or 4DEnVar. This agrees with the findings of Wang *et al.* (2013). A non-hybrid 4DEnVar system would be easier to maintain – it can be extended to handle additional model fields such as cloud without the need to build a static covariance model handling their correlation with other variables.

This work has not focussed on the computational efficiency of the methods, but some comments can be made based on our experience with the methods in our system. The main problems with hybrid-4DVar, which have driven us to investigate and others to adopt hybrid-4DEnVar instead, are (a) the design and maintenance of the linear model and its adjoint for an advanced forecast model and (b) the problems in the evaluation (in one form or another) of $\underline{\mathbf{M}}\mathbf{B}\underline{\mathbf{M}}^{\mathrm{T}}$ on parallel computers, given that the iterations and the forecast process are inherently sequential.

Solutions have been proposed to (a) by using automatic software generation and to (b) by Fisher and Gürol (2017), but they remain to be proven effective in global NWP. In contrast, in 4DEnVar the problem is the evaluation of terms like Equation 19; the initial reading and storage of all the $\underline{\mathbf{x}}'_{\mathrm{e},k}$ is a challenge for a large ensemble, especially if using methods as in this paper to improve $\mathbf{B}_{\text{ens}}$, which multiply the effective ensemble size and hence the costs. Each iteration needs a large number of transforms $\mathbf{U}^{\alpha}$ and Schur products. While all these can readily be parallelized, a good software design, minimizing communication costs, is essential.

Perhaps the deciding factor between the methods will be their suitability for extension to handle new science in DA for NWP, such as the precipitation-affected satellite radiances and radar reflectivity observations which are expected to give further improvements in weather forecasts. 4DVar relies on the representation of the processes being capable of linearization in **M**, whereas 4DEnVar can use values processed from an ensemble of nonlinear forecasts (Wang and Wang, 2017), but relies on the resulting distribution being near Gaussian.

ORCID

Andrew C. Lorenc https://orcid.org/0000-0001-7978-3594

REFERENCES

Bannister, R.N. (2017) A review of operational methods of variational and ensemble-variational data assimilation. *Quarterly Journal of the Royal Meteorological Society*, 143, 607–633. https://doi.org/10.1002/qj.2982.

Bloom, S.C., Takacs, L.L., Silva, A.M.D. and Ledvina, D. (1996) Data assimilation using incremental analysis updates. *Monthly Weather Review*, 124, 1256–1271.

Bocquet, M. and Carrassi, A. (2017) Four-dimensional ensemble variational data assimilation and the unstable subspace. *Tellus A*, 69(1). https://doi.org/10.1080/16000870.2017.1304504.

Bonavita, M., Hólm, E.V., Isaksen, L. and Fisher, M. (2016) The evolution of the ECMWF hybrid data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, 142, 287–303. https://doi.org/10.1002/qj.2652.

Bowler, N.E., Arribas, A., Mylne, K.R., Robertson, K.B. and Beare, S.E. (2008) The MOGREPS short-range ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society*, 134, 703–722. https://doi.org/10.1002/qj.234.

Bowler, N.E., Clayton, A.M., Jardak, M., Jermey, P.M., Lorenc, A.C., Wlasak, M.A., Barker, D.M., Inverarity, G.W. and Swinbank, R. (2017a) The effect of improved ensemble covariances on hybrid variational data assimilation. *Quarterly Journal of the Royal Meteorological Society*, 143, 785–797. https://doi.org/10.1002/qj.2964.

Bowler, N.E., Clayton, A.M., Jardak, M., Lee, E., Lorenc, A.C., Piccolo, C., Pring, S.R., Wlasak, M.A., Barker, D.M., Inverarity, G.W. and Swinbank, R. (2017b) Inflation and localization tests in the development of an ensemble of 4D-ensemble variational assimilations. *Quarterly Journal of the Royal Meteorological Society*, 143, 1280–1302. https://doi.org/10.1002/qj.3004.

Buehner, M. (2005) Ensemble-derived stationary and flow-dependent background-error covariances: evaluation in a quasi-operational NWP setting. *Quarterly Journal of the Royal Meteorological Society*, 131, 1013–1043. https://doi.org/10.1256/qj.04.15.

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

Buehner, M., Houtekamer, P.L., Charette, C., Mitchell, H.L. and He, B. (2010a) Intercomparison of variational data assimilation and the ensemble Kalman filter for global deterministic NWP. Part I: description and single-observation experiments. *Monthly Weather Review*, 138, 1550–1566. https://doi.org/10.1175/2009MWR3157.1.

Buehner, M., Houtekamer, P.L., Charette, C., Mitchell, H.L. and 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. *Monthly Weather Review*, 138, 1567–1586. https://doi.org/10.1175/2009MWR3158.1.

Buehner, M., McTaggart-Cowan, R., Beaulne, A., Charette, C., Garand, L., Heilliette, S., Lapalme, E., Laroche, S., Macpherson, S.R., Morneau, J. and Zadra, A. (2015) Implementation of deterministic weather forecasting systems based on ensemble-variational data assimilation at Environment Canada. Part I: the global system. *Monthly Weather Review*, 143(7), 2532–2559. https://doi.org/10.1175/MWR-D-14-00354.1.

Clayton, A.M., Lorenc, A.C. and Barker, D.M. (2013) Operational implementation of a hybrid ensemble/4D-Var global data assimilation system at the Met Office. *Quarterly Journal of the Royal Meteorological Society*, 139, 1445–1461. https://doi.org/10.1002/qj.2054.

Courtier, P., Thépaut, J.-N. and Hollingsworth, A. (1994) A strategy for operational implementation of 4D-Var, using an incremental approach. *Quarterly Journal of the Royal Meteorological Society*, 120, 1367–1387. https://doi.org/10.1002/qj.49712051912.

Desroziers, G., Camino, J.T. and Berre, L. (2014) 4DEnVar: link with 4D state formulation of variational assimilation and different possible implementations. *Quarterly Journal of the Royal Meteorological Society*, 140, 2097–2110. https://doi.org/10.1002/qj.2325.

Desroziers, G., Arbogast, E. and Berre, L. (2016) Improving spatial localization in 4DEnVar. *Quarterly Journal of the Royal Meteorological Society*, 142, 3171–3185. https://doi.org/10.1002/qj.2898.

Etherton, B.J. and Bishop, C.H. (2004) Resilience of hybrid ensemble/3DVAR analysis schemes to model error and ensemble covariance error. *Monthly Weather Review*, 132, 1065–1080. https://doi.org/10.1175/1520-0493(2004)132<1065:ROHDAS>2.0.CO;2.

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

Fisher, M. and Gürol, S. (2017) Parallelization in the time dimension of four-dimensional variational data assimilation. *Quarterly Journal of the Royal Meteorological Society*, 143, 1136–1147. https://doi.org/10.1002/qj.2997.

Fisher, M., Leutbecher, M. and Kelly, G.A. (2005) On the equivalence between Kalman smoothing and weak-constraint four-dimensional variational data assimilation. *Quarterly Journal of the Royal Meteorological Society*, 131, 3235–3246.

Flowerdew, J. and Bowler, N.E. (2011) Improving the use of observations to calibrate ensemble spread. *Quarterly Journal of the Royal Meteorological Society*, 137, 467–482. https://doi.org/10.1002/qj.744.

Gauthier, P. and Thépaut, J.-N. (2001) Impact of the digital filter as a weak constraint in the pre-operational 4DVAR assimilation system of Météo-France. *Monthly Weather Review*, 129, 2089–2102.

Houtekamer, P., Deng, X., Mitchell, H.L., Baek, S.J. and Gagnon, N. (2014) Higher resolution in an operational ensemble Kalman filter. *Monthly Weather Review*, 142(3), 1143–1162.

Ide, K., Courtier, P., Ghil, M. and Lorenc, A.C. (1997) Unified notation for data assimilation: operational, sequential and variational. *Journal of the Meteorological Society of Japan*, 75, 181–189.

Inverarity, G.W., Wlasak, M.A., Jardak, M. and Lorenc, A.C. (2018) *Deterministic data assimilation developments – July 2017*. Forecasting Research Technical Report 625. Met Office, Exeter, UK.

Kalnay, E., Li, H., Miyoshi, T., Yang, S.C. and Ballabrera-Poy, J. (2007) 4D-Var or ensemble Kalman filter? *Tellus A*, 59, 758–773. https://doi.org/10.1111/j.1600-0870.2007.00261.x.

Kepert, J. (2009) Covariance localization and balance in an ensemble Kalman filter. *Quarterly Journal of the Royal Meteorological Society*, 135, 1157–1176. https://doi.org/10.1002/qj.443.

Lei, L. and Whitaker, J.S. (2016) A four-dimensional incremental analysis update for the ensemble Kalman filter. *Monthly Weather Review*, 144(7), 2605–2621. https://doi.org/10.1175/MWR-D-15-0246.1.

Liu, C., Xiao, Q. and Wang, B. (2008) An ensemble-based four-dimensional variational data assimilation scheme. Part I: technical formulation and preliminary test. *Monthly Weather Review*, 136, 3363–3373. https://doi.org/10.1175/2008MWR2312.1.

Liu, C., Xiao, Q. and Wang, B. (2009) An ensemble-based four-dimensional variational data assimilation scheme. Part II: observing system simulation experiments with advanced research WRF (ARW). *Monthly Weather Review*, 137, 1687–1704. https://doi.org/10.1175/2008MWR2699.1.

Lorenc, A.C. (2003) The potential of the ensemble Kalman filter for NWP – a comparison with 4D-Var. *Quarterly Journal of the Royal Meteorological Society*, 129, 3183–3203. https://doi.org/10.1256/qj.02.132.

Lorenc, A.C. (2013) Recommended nomenclature for EnVar data assimilation methods. In *Research Activities in Atmospheric and Oceanic Modelling, WGNE*, http://bluebook.meteoinfo.ru/uploads/2013/chapters/BB_13_S1.pdf; accessed 2 August 2018.

Lorenc, A.C. (2017) Improving ensemble covariances in hybrid-variational data assimilation, without increasing ensemble size. *Quarterly Journal of the Royal Meteorological Society*, 143, 1062–1072. https://doi.org/10.1002/qj.2990.

Lorenc, A.C. and Rawlins, F. (2005) Why does 4D-Var beat 3D-Var? *Quarterly Journal of the Royal Meteorological Society*, 131, 3247–3257. https://doi.org/10.1002/qj.49711146703.

Lorenc, A.C., Ballard, S.P., Bell, R.S., Ingleby, N.B., Andrews, P.L.F., Barker, D.M., Bray, J.R., Clayton, A.M., Dalby, T., Li, D., Payne, T.J. and Saunders, F.W. (2000) The Met Office global three-dimensional variational data assimilation scheme. *Quarterly Journal of the Royal Meteorological Society*, 126, 2991–3012. https://doi.org/10.1002/qj.49712657002.

Lorenc, A.C., Bowler, N.E., Clayton, A.M., Fairbairn, D. and Pring, S.R. (2015) Comparison of hybrid-4DEnVar and hybrid-4DVar data assimilation methods for global NWP. *Monthly Weather Review*, 143, 212–229. https://doi.org/10.1175/MWR-D-14-00195.1.

Ménétrier, B., Montmerle, T., Michel, Y. and Berre, L. (2015a) Linear filtering of sample covariances for ensemble-based data assimilation. Part I: optimality criteria and application to variance filtering and covariance localization. *Monthly Weather Review*, 143(5), 1622–1643. https://doi.org/10.1175/MWR-D-14-00157.1.

Ménétrier, B., Montmerle, T., Michel, Y. and Berre, L. (2015b) Linear filtering of sample covariances for ensemble-based data assimilation. Part II: application to a convective-scale NWP model. *Monthly Weather Review*, 143(5), 1644–1664. https://doi.org/10.1175/MWR-D-14-00156.1.

Miyoshi, T., Kondo, K. and Imamura, T. (2014) 10240-member ensemble Kalman filtering with an intermediate AGCM. *Geophysical Research Letters*, 41, 5264–5271. https://doi.org/10.1002/2014GL060863.

Parrish, D.F. and Derber, J.C. (1992) The National Meteorological Center's spectral statistical-interpolation analysis system. *Monthly Weather Review*, 120, 1747–1763.

Payne, T. (2009) A methodology for forming components of the linear model in 4D-Var with application to the marine boundary layer. *Atmospheric Science Letters*, 10, 255–261. https://doi.org/10.1002/asl.234.

Payne, T.J. (2017) Rapid update cycling with delayed observations. *Tellus A*, 69(1). https://doi.org/10.1080/16000870.2017.1409061.

Rabier, F. (2005) Overview of global data assimilation developments in numerical weather prediction centres. *Quarterly Journal of the Royal Meteorological Society*, 131, 3215–3233. https://doi.org/10.1256/qj.05.129.

Rawlins, F., Ballard, S.P., Bovis, K.J., Clayton, A.M., Li, D., Inverarity, G.W., Lorenc, A.C. and Payne, T.J. (2007) The Met Office global four-dimensional variational data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, 133, 347–362. https://doi.org/10.1002/qj.32.

Raynaud, L., Berre, L. and Desroziers, G. (2011) An extended specification of flow-dependent background-error variances in the Météo-France global 4D-Var system. *Quarterly Journal of the Royal Meteorological Society*, 137, 607–619. https://doi.org/10.1002/qj.795.

Stiller, O. (2009) Efficient moist physics schemes for data assimilation. II: deep convection. *Quarterly Journal of the Royal Meteorological Society*, 135, 721–738.

Stiller, O. and Ballard, S.P. (2009) Efficient moist physics schemes for data assimilation. I: large-scale clouds and condensation. *Quarterly Journal of the Royal Meteorological Society*, 135, 707–720.

Thépaut, J.-N., Courtier, P., Belaud, G. and Lemaître, G. (1996) Dynamical structure functions in a four-dimensional variational assimilation: a case-study. *Quarterly Journal of the Royal Meteorological Society*, 122, 535–561.

Trevisan, A., D'Isidoro, M. and Talagrand, O. (2010) Four-dimensional variational assimilation in the unstable subspace and the optimal subspace dimension. *Quarterly Journal of the Royal Meteorological Society*, 136, 487–496. https://doi.org/10.1002/qj.571.

Wang, X., Parrish, D., Kleist, D. and Whitaker, J.S. (2013) GSI 3DVar-based ensemble-variational hybrid data assimilation for NCEP global forecast system: single resolution experiments. *Monthly Weather Review*, 141(11), 4098–4117. https://doi.org/10.1175/MWR-D-12-00141.1.

Wang, Y. and Wang, X. (2017) Direct assimilation of radar reflectivity without tangent linear and adjoint of the nonlinear observation operator in the GSI-based EnVar system: methodology and experiment with the 8 May 2003 Oklahoma City tornadic supercell. *Monthly Weather Review*, 145(4), 1447–1471. https://doi.org/10.1175/MWR-D-16-0231.1.

Wlasak, M. and Cullen, M.J.P. (2014) Modelling static 3D spatial background-error covariances – the effect of vertical and horizontal transform order. *Advances in Science and Research*, 11(1), 63–67. https://doi.org/10.5194/asr-11-63-2014.

How to cite this article: Lorenc AC, Jardak M. A comparison of hybrid variational data assimilation methods for global NWP. *Q J R Meteorol Soc.* 2018;144:2748–2760. https://doi.org/10.1002/qj.3401