TECHNICAL MEMORANDUM

A Strategy for Data Assimilation

Massimo Bonavita, Yannick Trémolet, Elias Holm, Simon T. K. Lang, Marcin Chrust, Marta Janisková, Philippe Lopez, Patrick Laloyaux, Patricia de Rosnay, Mike Fisher, Mats Hamrud, Stephen English

Research Department & Forecast Department

April 2017

Series: ECMWF Technical Memoranda

A full list of ECMWF Publications can be found on our web site under:
http://www.ecmwf.int/publications/

Contact: library@ecmwf.int

Abstract

This paper reviews progress made in Data Assimilation methodology at ECMWF since the 2011 SAC Data Assimilation Vision document (ECMWF/SAC/40(11)7). It then describes ECMWF's current overall strategy for Data Assimilation. Recent research and development activities in data assimilation methodology have been centred on the introduction of an ensemble data assimilation component for error estimation and cycling and the development of a software framework for more flexible code design through the Object Oriented Prediction System (OOPS). In addition, investigation of ensemble Kalman filter methods, the use of weak constraint 4D-Var and the sensitivity to assimilation window length are active areas of research. A new aspect discussed in this paper relates to the challenges posed by the goal to develop a data assimilation system able to provide physically consistent initial estimates for all the modelled Earth system components.

1   Introduction

This paper is based on a Special Topic Paper prepared for the Scientific Advisory Committee meeting held at ECMWF in October 2016 (ECMWF/SAC/45(16)9). This follows ECMWF's previous paper on data assimilation strategy, Data Assimilation Vision 2014-2018 (ECMWF/SAC/40(11)7), which included ambitious plans for a new flexible code design through the Object Oriented Prediction System (OOPS); the extension of variational methods to provide flow dependent background errors and initial conditions for ensemble forecasting; investigation of ensemble Kalman filters; the use of weak constraint and longer windows in 4D-Var; ideas for further code-efficiency and more optimal and frequent operational analysis updates through overlapping windows. There has been progress in all areas, in particular ensemble data assimilation (Bonavita, Hólm, Isaksen and Fisher 2016). In addition significant progress has been made in coupled data assimilation (Laloyaux, Balmaseda, Dee, Mogensen and Janssen 2016).

In this paper we describe progress in implementing the previous strategy and present an updated strategy for data assimilation based on this experience, progress elsewhere and the goals of the ECMWF Strategy 2016-2025. The main elements are ensemble data assimilation, coupled data assimilation, the 4D-Var algorithm itself, an improved framework for the assimilation code, algorithmic efficiency, parallelism improvements, and fuller use of all available observations.

OOPS will remain a central element for the implementation of the strategy, opening the door for the scientific evolution of the data assimilation algorithms, providing a common infrastructure for all components of the Earth system, and facilitating collaborations in data assimilation. These collaborations will be critical to the successful delivery of the strategy.

2   The ECMWF context

2.1   The high resolution analysis cycle

Data Assimilation is a fundamental component of any numerical weather prediction (NWP) system and has been instrumental to the sustained improvement in the skill of ECMWF forecasts over the past twenty years. During this period, the mainstay of data assimilation at ECMWF has been the 4D-Var algorithm, whose development started in the late eighties (Le Dimet and Talagrand 1986; Courtier,

Thépaut and Hollingsworth 1994) and whose first operational implementation was realised at ECMWF in the late nineties (Rabier, Järvinen, Klinker, Mahfouf and Simmons 2000).

The success of the standard 4D-Var assimilation approach has been based, among other things, on the ability of the variational algorithm to make effective use of the ever increasing quantity of asynchronous, satellite-based observations in a manner consistent with the model dynamics; on its capacity to account for weak nonlinearities in the model evolution and the observation operators; and on the ability to perform a global analysis, i.e. to avoid using localization techniques to restrict the influence of observations which are not close (in space and in time) to the analysed grid point.

Another area of strength for 4D-Var is its ability to incorporate additional terms in the cost function. This has enabled the adoption of variational bias correction of observations (Dee 2004), variational control of gravity waves via an additional constraint based on a Jc-Digital Filter Initialisation (Jc-DFI, Gauthier and Thépaut 2001) and correction of model bias (Trémolet 2006; ECMWF/SAC/45(16)8 2016).

In its current operational use, 4D-Var is run with three outer iterations. All the trajectories run at the same resolution as the forecast, TCo1279. The three inner iterations are performed at resolutions of TL255-319-399 for an increasingly refined analysis.

2.2   The error cycling component: EDA

The ensemble DA assimilation system (Berre, Stefanescu and Belo Pereira 2006; Belo Pereira and Berre 2006; Isaksen, Bonavita, Buizza, Fisher, Haseler, Leutbecher and Raynaud 2010) used at ECMWF and Météo-France is a Monte Carlo simulation of the reference assimilation system (Bonavita, Isaksen and Hólm 2012) accounting for observations, forcing and model error sources. It is capable of handling non-Gaussian posterior Probability Density Function (pdf) provided the ensemble size is sufficiently large. Its maintenance cost is low because there is no need to support a separate data assimilation system for the ensemble component.

The EDA is an ensemble of 4D-Var analysis cycles. Therefore it is simple to propagate new developments from the high resolution 4D-Var to the EDA. The model, the observations and the boundary conditions (sea surface temperature) are perturbed according to their expected uncertainties. Starting with the preliminary experiments of Kucukkaraca and Fisher (2006), EDA information has been stepwise introduced into the background error variances and the wavelet background error correlation model, ensuring a smooth, full rank B model at all times (Fisher 2007; Isaksen, Fisher and Berner 2007; Bonavita, Raynaud and Isaksen 2011; Bonavita, Isaksen and Hólm 2012; Bonavita, Hólm, Isaksen and Fisher 2016). The current EDA consists of 25 members at the same TCo639 resolution as the ECMWF Ensemble Forecast system (ENS), and with the same L137 vertical levels as the TCo1279 high resolution deterministic analysis. The resolution of the two inner loops of the EDA have been recently increased from TL95-TL159 to TL191-191, but are still reduced with respect to the high resolution analysis, which are TL255-TL319-TL399. The EDA forecasts are qualitatively similar to the full resolution model, capturing small scale systems such as tropical cyclones almost as well. The recent resolution increase improved the statistical reliability of the EDA. The EDA variances are still globally inflated by a factor 1.3 before being used in 4D-Var due to sources of errors not represented in the EDA.

The EDA calculates variances of the day from the spread of the 25 members' background forecasts and covariance matrices from a mix of 400 samples spread through the year and 200 samples from the current cycle obtained from 8 hourly time steps centred on the background time. The resolution of the EDA variances is TL399, which matches the highest inner loop of 4D-Var, and the variances are signal-to-noise filtered by a non parametric wavelet-based filter (Bonavita, Isaksen and Hólm 2012) which is spatially heterogeneous and also eliminates small scale ringing artefacts in intense systems like tropical cyclones. This configuration of the EDA provides an affordable way to produce flow dependent covariance estimates. Further weighting towards the covariances of the day for the smaller scales is achieved because wavebands above T63 up to T399 keep constant horizontal resolution for the vertical correlation matrices (one every 250km) but have more points per matrix at higher wavenumbers. Up to T63, 30% of the covariance comes from the currently valid samples and this increases gradually up to 93% at T399, compared with 100% for the variances.

Two further developments which are being pursued are the extension of the use of EDA error estimates to humidity variances (currently a function of background relative humidity and level) and later cloud variable variances, and the introduction of coordinate transforms to add horizontal anisotropy to the wavelet covariance model (see e. g. Desroziers 1997; Piccolo and Cullen 2012; Michel 2013a, b; Legrand and Michel 2014; Pannekoucke, Emili and Thual 2014). An isotropizing coordinate transform based on the sampled error correlation between neighbouring points redistributes the field so that the correlations are approximately isotropic in the transformed coordinate, where the diagonal wavelet model is a good approximation. The correlation matrices can be averaged over a larger area to increase statistical robustness, with the small scale correlations and local anisotropy and heterogeneity moved out to the grid transformation which locally rotates and dilates the grid.

2.3   Ensemble initialization

In the current operational configuration of the ENS, perturbations are added to the high resolution analysis to provide initial conditions for the perturbed ensemble members. The perturbations are generated from the perturbed EDA members with respect to the EDA mean and singular vectors (Buizza, Leutbecher and Isaksen 2008; see Leutbecher and Palmer 2008 for details of the singular vector perturbations). The perturbations have a plus-minus symmetry, i.e. the 25 EDA perturbations are distributed between the 50 ENS members in such a way that there are always two of the ENS members that share the same initial perturbation with the sign of the perturbation reversed.

The EDA perturbations are generated from 6h-forecasts from the perturbed EDA members. The perturbations are then re-centred on the high resolution analysis. The reasons for re-centring are twofold: the EDA members are run at lower resolution and with less outer loops than the high resolution 4D-Var and furthermore, there is no up-to-date EDA available at the time the ENS starts in the operational schedule. Re-centring is further discussed in Section 4.1.

2.4   OOPS

The IFS software has become complex, making it difficult to introduce and test new approaches. This complexity affects all parts of the system. For example observations require sophisticated observation operators, bias correction, and handling of correlated errors; model errors need to be estimated and accounted for; the tangent linear and adjoint models are increasingly sophisticated; the background error

covariance model relies on wavelets and nonlinear change of variables. The Earth System approach also increases this complexity, with more system components and the coupling between them. At the same time, computer hardware is also becoming more challenging to use efficiently, hence ECMWF's Scalability programme, and energy efficiency is increasingly a concern. Further to this ECMWF's Strategy relies on collaboration: more people and organisations working together with the code also generates a requirement for more flexible code. These issues are not specific to ECMWF; many NWP centres are examining how to modernise their code infrastructure for similar reasons. At ECMWF our project addressing this is called the Object Oriented Prediction System (OOPS). OOPS is part of the Scalability programme and is a collaborative project with Member States, in particular Météo-France. This particularly affects data assimilation because it brings together most if not all parts of the forecasting system.

Despite the complexity, all data assimilation methods manipulate the same limited number of entities: states, observations, increments, covariance matrices, models and observation operators. Using these entities as abstract building blocks, it is possible to develop data assimilation algorithms and other high level applications. These building blocks can be implemented for different models. Although the components above form its basic structure, an operational forecasting system also should be reliable, efficient, readable and flexible. The key behind achieving this is separation of concerns: all aspects exist but in different parts of the code and scientists focus on one aspect at a time. This is well recognised in the software industry where the tools for implementing such a structure are called generic and object oriented programming. OOPS was developed on this basis.

OOPS is implemented in C++ and the entities described above are the main classes in the code. Each class has well defined responsibilities: the State knows its values at any location, the Model advances the State in time, and the observation operator computes observation equivalents given the values of the State at the appropriate locations. For this, interfaces are the most important aspect of the design, which is very different from working practices with most existing FORTRAN codes.

Differently from other attempts at designing generic data assimilation systems (Anderson, Hoar, Raeder, Liu, Collins, Thorn and Avellano 2009; Chapelle, Fragu, Mallet and Moireau 2013), OOPS doesn't access the internal details of any data structures or computations: it only passes addresses or pointers between appropriate methods. The reason behind this choice is that it is impossible to cater for all possible data structures and distributions of data across processors. This should be dealt with at another level. For example, another project at ECMWF is developing Atlas, a software package to handle grids and interpolations. This addresses separation of concerns at project level.

Another important aspect of OOPS is that it includes testing. The design by classes facilitates the implementation of unit tests which are run automatically when changes are lodged in the common source code repository and can be run on demand for code in local repositories. These tests are for internal consistency and technical checks for known configurations. They are not intended for meteorological evaluation but should detect as many bugs as possible very early in development stages. The fact that OOPS runs 4D-Var as a single executable will also increase its efficiency by reducing the start-up times (the current code requires seven executables and a large amount of I/O to pass data between them).

OOPS has been developed and tested with two independent but very simple models. It has already been useful to validate scientific ideas, in particular related to weak constraint 4D-Var and the saddle point

algorithm. The refactoring of the two largest components of ECMWF's forecasting systems, the IFS and NEMOVAR, is progressing. Most of the work and resources are spent creating self-contained units of code that can be called through proper interfaces from the OOPS layer while still functioning in their current environment.

Because it imposes separation between the data assimilation layer and model specific aspects, OOPS is a good collaboration platform between different communities. It has been an opportunity to strengthen collaborations with Météo-France, HIRLAM and ALADIN for the IFS; and with the Met Office, CERFACS and INRIA for NEMOVAR. Météo-France is using OOPS to develop and evaluate the 4D-En-Var algorithm, so OOPS provides a framework for sharing developments. An open source version of OOPS with simplified models will be released. It should facilitate the uptake of data assimilation research in the academic community and in the future should be an attractive platform to attract collaborations and external projects, with simplified or operational models. Plans for similar systems, partly inspired by OOPS, are being discussed in the United States bringing together operational and academic data assimilation communities which could create further opportunities for collaborations.

3   Review of alternative approaches

Several different ensemble, variational and hybrid data assimilation methods are in use or at various stages of development in other NWP centres. In the following we give a brief overview of the main alternative approaches and how they relate to the current ECMWF data assimilation system.

3.1   The EnKF

The EnKF has been originally developed as a Monte Carlo approximation to the Kalman Filter (Evensen 1994; Burgers, van Leeuwen and Evensen 1998; Houtekamer and Mitchell 1998). This implementation of the EnKF, known as perturbed observation or stochastic EnKF, is currently run at Environment Canada to provide initial conditions for their ensemble prediction system. It is based on the concept of simulating all possible sources of uncertainty in the data assimilation system. This implies running an ensemble of data assimilation systems where random perturbations drawn from their expected error statistics are added to the observations, the model and the boundary conditions. Burgers, van Leeuwen and Evensen (1998) demonstrate that the sampled analysis error covariances from the perturbed observation EnKF asymptotically match the values computed by the Kalman filter update equations. Another possibility is to compute a Kalman filter update of the ensemble mean and the analysis covariance and then construct the ensemble members from a 'square root' of the sampled analysis covariance (EnSRF: Anderson 2001; Whitaker and Hamill 2002; Tippett, Anderson. Bishop, Hamill and Whitaker 2003, for a review). This version of EnKF is run operationally at NCEP. In order to be computationally efficient, the EnSRF assimilates observations sequentially, thus implicitly assuming their errors to be statistically uncorrelated. This assumption is not needed in the more recent version of the EnKF known as Local Ensemble Transform Kalman Filter (LETKF: Hunt, Kostelich and Szunyogh 2007), where the analysis is computed independently for each grid point using all the observations in a predefined local volume. An earlier variant of this type of EnKF is currently run operationally at the Met Office for the initialization of the ensemble forecast system. As both the EnSRF and the LETKF do not need to perturb the observations they are called deterministic EnKF. Whitaker and Hamill (2002) and Bowler, Flowerdew and Pring (2013) have suggested that deterministic EnKF is more efficient than its stochastic counterpart, i.e., it produces more accurate analyses for a given ensemble size. It is also

computationally cheaper to run than a perturbed observation EnKF as it requires the computation of only one analysis update for the ensemble mean and the ensemble perturbations. This has motivated the choice to implement versions of the EnSRF and LETKF at ECMWF (Hamrud, Bonavita and Isaksen 2015) for research evaluation. A comprehensive review of EnKF is given by Houtekamer and Zhang (2016).

Results described below have been obtained with the LETKF version of EnKF, due to its favourable scalability properties (Yashiro, Terasaki, Miyoshi and Tomita 2016) for assimilation systems ingesting large observational datasets ($O(10^7)$ in the ECMWF case). As with all ensemble data assimilation systems, in the forecast stage the scalability limit is set by the scalability of the individual forecast members. The LETKF analysis itself is local; in principle the limit of independent pieces of work is the size of the analysis state. As the EDA can be seen as a variational implementation of a perturbed observation EnKF, it shares the favourable scalability properties of the standard EnKF for the forecast stage. The only difference in scalability comes then from the analysis update step, which in the EDA is based on strong constraint 4D-Var. Due to the sequential nature of the variational minimization and the relatively low resolution at which the minimizations are performed the scalability properties of each EDA 4D-Var are less favourable with the current algorithm (see ECMWF/SAC/40(11)7 for a discussion), though several developments have the potential to improve the scalability characteristics of the EDA (section 4.1). Note however that the full EDA system scales perfectly with ensemble size as each EDA member is independent.

Computational cost is also an important consideration for an operational system. Improvements to the computational cost (and also the skill) of the EDA appear to be possible following the mean-perturbation approach of Lorenc, Jardak, Payne, Bowler and Wlasak (2016) (described in section 3.2) and also using information from an initial minimization of the control (unperturbed) EDA member to speed up the convergence of the perturbed EDA members (sections 4.1.4 and 4.2.1). An EnKF system can also cut costs compared to the current EDA configuration mainly due to the fact that the Kalman Gain needs only be computed once in the EnKF and not separately for each ensemble member as in the EDA.

The accuracy of the control state of a 100 member EnKF and a 25 member EDA, is comparable for tropospheric scores, while the EDA control forecast is more skilful in the stratosphere (Bonavita, Hamrud and Isaksen, 2015). It is likely that the vertical localization with a 100 member ensemble limits the information that the EnKF (and ensemble-variational methods such as 4D-En-Var) can extract from non-local observations. Improved localization is an active area of research (Flowerdew 2015; Ménétrier and Auligné 2015; Lei and Whitaker 2015). Whitaker and Lei (2016) have suggested that an ensemble size of dimension $O(1000)$ would allow vertical localization to be switched off completely for the current atmospheric observing system. Note that in the case of no vertical localization the LETKF would compute analysis updates for whole grid columns instead of grid points, which would significantly improve efficiency and balance characteristics of the ensuing analysis.

It is worth noting that two of the main arguments that convinced ECMWF to abandon its Optimal Interpolation (OI) analysis system in favour of 3D-Var were the variational method's superiority in handling radiance data and in the global nature of its analysis that avoided partitioning observations into local boxes (Andersson, Pailleux, Thépaut, Eyre, McNally, Kelly, and Courtier, 1994; Andersson, Haseler, Undén, Courtier, Kelly, Vasiljevic, Brankovic, Gaffard, Hollingsworth., Jakob, Janssen,

Klinker, Lanzinger, Miller, Rabier, Simmons, Strauss, Viterbo, Cardinali, and Thépaut 1998). Although the EnKF is undoubtedly a significant improvement over OI, the arguments put forward for adopting variational data assimilation remain valid today.

3.2   4D-En-Var

Ensemble-Variational methods (En-Var; Liu, Xiao and Wang 2008) aim at combining the advantages of ensemble Kalman filters and variational methods. The En-Var approach solves the assimilation problem in the sub-space of ensemble perturbations as EnKF, but through the minimization of a cost function. This results in the possibility of re-using a large part of the standard 4D-Var code base and low maintenance costs when 4D-Var is already available. En-Var shares many of the benefits of the existing 4D-Var such as bias correction, correlated observation error handling and Jc-DFI.

In 4D-Var, the time evolution of the background error covariances inside the assimilation window is implicitly generated by the application of the tangent linear and adjoint models to the analysis increments while in the En-Var algorithm localised ensemble perturbations are used to sample the B matrix throughout the assimilation window. Thus En-Var methods do not require tangent linear and adjoint versions of the forecast model and are also cheaper than 4D-Var but, on the other hand, are affected by sampling and localization issues which arise from the limited dimension of the ensemble, similarly to the EnKF.

In its initial implementation, En-Var is a deterministic analysis algorithm, i.e. it only provides an update for the mode of the analysis ensemble. In order to obtain a sample of the full analysis ensemble, ensembles of perturbed observations 4D-En-Vars have been developed (En-4D-En-Var). This algorithm uses the principle of self-exclusion, where the gain computation for each member uses the perturbations from all the other members but not from itself. Full cycling of errors is then possible, generating flow dependent background error perturbations that can be used in the high resolution deterministic 4D-Var with positive impact (Bowler, Clayton, Jardak, Lee, Lorenc, Piccolo, Pring, Wlasak, Barker, Inverarity and Swinbank 2016a).

In order to reduce further the computational cost of En-4D-En-Var, a modification to the ensemble generation mechanism has been recently proposed under the name of "mean-perturbation" approach (Lorenc, Jardak, Payne, Bowler and Wlasak 2016). The idea is to compute separately the analysis update for the ensemble mean and for the perturbations from the mean. The computational savings arise from employing a simplified, cheaper configuration with respect to that used in the minimization of the ensemble mean analysis (e.g., reduced number of iterations, and/or 3D-Var cost function, and/or reduced outer loop resolution, etc.) in the minimization of the cost function for the perturbations. Although developed in the En-Var context, a similar approach to the "mean-perturbation" method can be applied in the EDA context as shown in section 4.2.1.

The En-Var approach has been adopted for the high resolution DA system at Environment Canada (Buehner, McTaggart-Cowan, Beaulne, Charette, Garand, Heilliette, Lapalme, Laroche, Macpherson, Morneau and Zadra 2015) and NCEP (NCEP, 2016). The Met Office is planning to replace their EnKF by En-4D-En-Var in the near future for their global ensemble data assimilation system, although they have shown that for the high resolution analysis, 4D-Var is better than 4D-En-Var; Bowler, Clayton, Jardak, Lee, Lorenc, Piccolo, Pring, Wlasak, Barker, Inverarity and Swinbank 2016b). Given its

potential and synergy with the existing 4D-Var, it is important for ECMWF to develop an En-Var capability. The main 4D-En-Var algorithm has already been implemented in OOPS for toy models. For use with the IFS, it still requires a localisation matrix in model space. This is currently being developed at Météo-France and we expect it will be a product of the IFS/Arpege collaboration and of the OOPS project.

3.3   Extended Control Variable

With regards to the question of how to incorporate ensemble information in the high resolution analysis, the ECMWF approach so far has been to model the background error covariance matrix. This model is then fitted to samples of background errors that depend on the meteorological situation of the day and are generated by the EDA system.

An alternative to fitting B on the samples of the day is to reformulate it as a combination of a fixed climatological term and a flow dependent component generated from the background perturbations from the ensemble. In that case, the analysis increment in model space can be expressed as the sum of the increment in the sub-space spanned by the climatological B and a linear combination (with coefficients α) of the sample of forecast perturbations from the ensemble (this method is often called the α, or extended, control variable, Lorenc, 2003). This is equivalent to a linear combination of the climatological and flow-dependent error covariance matrices (Wang, Snyder and Hamill 2007).

In practice, because of the limited size of the ensemble, its contributions are localised. This is achieved by allowing the coefficients α to slowly vary geographically. The values of the α control variable are constrained by an additional term in the cost function of the form $J = 1/2\ \alpha^T\ C_{loc}^{-1}\ \alpha$, where the C_{loc} matrix is an empirical covariance matrix that controls the spatial variation of the α control variable and thus the amount of localization applied to the ensemble perturbations. This type of hybrid assimilation scheme has been implemented in the Weather Research and Forecast system (WRF, Wang, Barker, Snyder and Hamill 2008) and is currently run operationally at the Met Office (Clayton, Lorenc and Barker 2013).

Positive results were obtained at ECMWF by averaging one analysis issued from 4D-Var with another more localised one issued from EnKF (see section 4.3) but with some drawbacks, in particular related to initialisation and spin-up. The α control variable approach combines the global and full rank aspects of 4D-Var with a localised and flow dependent component from the ensemble, but optimised together in a consistent manner, thus in principle avoiding the drawbacks mentioned above. Because it is performed in the 4D-Var context, this combined analysis would also benefit from additional terms in the cost function such as Jc-DFI.

The α control variable extension to 4D-Var is coded in the OOPS framework. However, for use with the IFS, it requires a localisation matrix in model space. The localisation being developed in the context of 4D-En-Var at Météo-France could be used for this purpose. Although quite different in concept, there are synergies between 4D-En-Var and the α control variable in 4D-Var, both from a technical point of view and from a scientific point of view related to the understanding of model space localisation. Developments will continue in the same framework to build on the synergies as much as possible.

3.4   Particle Filters

The aim of particle filters is to estimate the state pdf by a collection of states (particles) which are generated by parallel model runs. The simplest form of the particle filter starts from a collection of background states, each having a weight according to its probability, then integrates the states to the next analysis time and finally updates the weights for each state according to the likelihood of the observations given that state using Bayes' theorem (see van Leeuwen 2015 for a review). The main claim for the particle filter is that it can handle arbitrary pdf for the state, thus accommodating nonlinearities in the model evolution and observation operators without the need to explicitly handle state covariances. Several developments are ongoing to make particle filters useful in high-dimensional systems such as NWP. Only a few recent developments are given here, with further information found in the references. A problem with the simplest particle filter is that the weights will quickly converge on one state for moderate number of particles, with all others having essentially zero weight. Therefore resampling is introduced, with more likely states replicated and less likely states removed. However, even this is not enough to prevent filter collapse, and several techniques have been introduced to improve the sampling (van Leeuwen 2015). One technique is localization, which reduces the number of observations affecting each point (e. g. Reich 2013; Poterjoy and Anderson 2016; Robert and Kuensch 2016). With localization Poterjoy and Andersson (2016) were able to use 25 particles to analyse a coarse resolution Global Circulation Model. The number of members needed by a particle filter is an important factor in its efficiency and objective criteria for how many particles are needed for a given system are given by Slivinski and Snyder (2016) for certain classes of particle filters (these criteria do however not carry over automatically to other filters like e. g. Ades and van Leeuwen 2015). With localization, the weight of particles varies in space and one particle may have large weight in one area and low in an adjacent area. This can cause unphysical behaviour, and to counter this Reich (2013) introduced additional constraint that gives smooth variation of the weights by an optimization technique called optimal transport. In general all filters that use localization add measures to ensure smooth updates, which complicates the update, as in e. g. Poterjoy and Anderson (2016). An issue with all explicitly or implicitly localized filters (Ades and van Leeuwen 2015, being example of the latter due to their particular treatment of model error) is that dynamical balances may not be preserved, and this has been investigated by Ades and van Leeuwen (2015). Another main technique used in particle filters is to introduce a proposal transition density which depends on the past state, current observations and model error parameterization (van Leeuwen 2015). The right hand side of Bayes' theorem is multiplied by the proposal density in both the numerator and the denominator, and the proposal density is thus arbitrary. However, by choosing the proposal density wisely, particles can be pre-guided to the most likely states. The proposal density in the numerator now becomes what is sampled and the proposal density in the denominator is absorbed into the weight calculation where it compensates for not sampling the original model, and sometimes is chosen to simplify the weight calculation. In principle, sampling a proposal density instead of the original model corresponds to a pre-analysis which can be performed by any analysis method, from nudging (Haslehner, Jancic and Craig 2016; Browne and van Leeuwen 2015), to ensemble Kalman filter (Robert and Kuensch 2016) and even 4D-Var, where a simplified EDA could be used. Gradually larger geophysical systems are being analysed with moderate number of particles and Browne and van Leeuwen (2015) analysed a coupled atmosphere-ocean model with over 2 million variables and ca. 27 thousand independent sea surface temperature observations per time step with only 32 members. The number of affordable particles for high-dimensional systems like NWP puts a limit on how much of the nonlinearity of the analysis pdf can really be captured by particle filters. One way to

address this is by combining particles at different resolution as done by Gregory, Cotter and Reich (2016) which combine the so-called multilevel Monte Carlo method with the particle filter of Reich (2013). Further research on making particle filters efficient in high-dimensions include hybridization with ensemble Kalman filters (e. g. Robert and Kuensch 2016) and better proposal densities (e. g. Zhu, van Leeuwen and Amezcua 2016). As the particle filters become more efficient for high-dimensional systems, a tendency is seen to include more of the currently used assimilation algorithms such as EnKF or variational methods with their covariance modelling into the proposal densities. Therefore developments aimed at improving the efficiency of particle filters will be of relevance to other methods as well and it is anticipated that in the coming years ideas developed in the context of particle filters could be taken and applied at ECMWF. At this point it is not yet anticipated that particle filters themselves are an operational development path, but this is a rapidly developing field and it will be important to keep this under review in the coming years.

3.5   EVIL

The Ensemble-Variational Integrated Localised (EVIL, Auligné, Ménétrier, Lorenc and Buehner 2016) Data Assimilation is similar in concept to the reduced-rank Kalman Filter developed at ECMWF in the late nineties (Fisher 1998; Fisher and Andersson 2001). The idea is to make use of the approximate eigenvectors of the Hessian matrix of the cost function computed in the high resolution analysis to construct a linear transformation that maps the background ensemble perturbations into analysis ensemble perturbations which are then added to the high resolution analysis to provide an updated analysis ensemble. These eigenvectors are a free by-product of the combined Lanczos/conjugate gradient algorithm used in the minimization (Fisher and Andersson 2001). This method has two main advantages. First, the posterior ensemble perturbations are consistent with the diagnosed errors of the high resolution variational analysis. Second, there is only one system to maintain, the variational one. Three versions are considered which are similar to the standard perturbed EnKF (S-EVIL), the deterministic EnKF (D-EVIL) and a resampling version (R-EVIL). The resampling version can generate any number of consistent initial states for ensemble forecasting, independent of the number of prior perturbations, through a transform of normally distributed random vectors with zero mean and unit variance. In the extreme case only a static B matrix is needed, but in general the advantage of EVIL is that the posterior perturbations get information from full B modelling that includes both ensemble and climatological components. Another difference between the versions is that the observation operators need to be evaluated for each member in S-EVIL, in addition to the evaluation in the deterministic analysis in all three versions. The main research issue for EVIL application in NWP is the number of eigenvectors needed for high-quality update of the background to analysis perturbations, because more updates increase the cost of the method through further iterations of the deterministic minimization, or through the calculation of the eigenvectors in another way. The number of eigenvectors needed may be thousands using a realistic observing network (Auligné, Ménétrier, Lorenc and Buehner 2016; Buehner, McTaggart-Cowan and Heilliette 2016a, b). Further development of this methods will need to address more efficient eigenvector calculations, and ideas include using the Block-Lanczos algorithm to perform EVIL on multiple perturbed right-hand sides to increase the number of eigenvectors, which is a development similar to future plans to increase the efficiency of the EDA.

For these reasons at the moment ECMWF does not plan to develop EVIL, but as with particle filters this is an area where developments need to be followed, and these developments are relevant to the EDA.

4   The ECMWF hybrid 4D-Var: research perspectives and science plan

4.1   4D-Var

4.1.1   Weak constraint 4D-Var

The theoretical formulation of 4D-Var comprises estimating the four dimensional state of the atmosphere (or system of interest) over the length of the assimilation window. The perfect model assumption reduces the control variable to the initial condition at the beginning of the assimilation window, relying on the model to determine the solution at other times within the window. This assumption made 4D-Var possible in the late 90's. As other aspects of the system have progressed, model error becomes more relevant to properly account for. This is the case in ensemble prediction and assimilation (Leutbecher et al., 2016). More than compensating for model error, weak constraint 4D-Var is currently the only assimilation algorithm that can provide an estimate of model error directly constrained by observations. It thus provides information that can "close the loop" on model uncertainty.

Theory also shows that under appropriate assumptions weak constraint 4D-Var is equivalent to a full rank Kalman smoother (Fisher, Leutbecher and Kelly 2005; Ménard and Daley 1996). Using a four dimensional control variable, weak constraint 4D-Var is in effect splitting the assimilation window in sub-windows between the times when the control variable is defined. An implementation of long window weak-constraint 4D-Var with overlapping assimilation windows that move forward in time by one sub-window at a time is an approximation of weak constraint 4D-Var with an assimilation window that extends (almost) indefinitely in the past (Trémolet 2006). In this case, the background error covariance matrix does not need to be specified as it evolves implicitly within the assimilation window. Only the observation and model error covariance matrices are required.

However, although weak constraint 4D-Var has attracted attention recently (Cullen 2010, 2013; Todling 2015; Bishop 2016; Bowler 2016), the model error covariance matrix, Q, remains largely unknown. Q can be estimated using observations, through the data assimilation system. The most promising line of research in that direction seems to be the extension of the work by Todling (2015) in the context of lagged smoothers to the variational framework. However, Q should not depend on observations (unlike background errors that do depend on previous observations). Based on that, it could be preferable to estimate Q without relying on data assimilation to avoid potential aliasing. This is the direction that was chosen when the estimation of systematic model error was re-introduced in Cycle 43r1. The estimation of Q is derived from an ensemble where the model is perturbed but not the initial conditions (Leutbecher et al., 2016). So far this approach has been used only for estimating the slowly evolving component of model error, this research will be pursued in the context of a fully four dimensional control variable. In the longer term, since weak constraint 4D-Var is the only method that produces an estimate of model error, the potential for use of an ensemble of weak constraint 4D-Vars for an online estimation of model error covariances might be investigated.

4.1.2   Overlapping windows, quasi-continuous DA

A full implementation of weak-constraint 4D-Var is not achievable yet but the principle of overlapping assimilation windows can be used within strong constraint 4D-Var. For the high resolution assimilation

cycle the current operational configuration consists of cycling long cut-off 12-hour 4D-Var analyses, from which simplified "early delivery" 6-hour 4D-Var analyses are also generated. A promising alternative configuration is to run four overlapping 12-hour 4D-Var analyses a day, with relatively tight observation cut-off times, in order to guarantee the timely availability of updated initial conditions at synoptic times. Observations that do not arrive in time for the current analysis update (around 10% with present "early-delivery" observation cut-off times) would be picked up by the subsequent analysis, guaranteeing a nearly complete use of available observations. The possibility to use in 4D-Var an initial linearisation state from the previous overlapped analysis window, which has seen a subset of the batch of observations used in the current analysis window, is also expected to improve the convergence properties of the minimization. The choice of an initial linearization trajectory primarily affects the rate of convergence of 4D-Var, not the solution, in weakly non-linear data assimilation. Thus this does not introduce undesirable correlations between background and observations.

In the future, this system can be extended, keeping the 12-hour assimilation windows, but running them much more frequently (this is conceptually similar to the quasi-continuous DA discussed by Järvinen, Thépaut, and Courtier 1996; Pires, Vautard and Talagrand 1996). For example if we update every hour the assimilation window can be extended all the way to the current time and all observations that have arrived in the last hour can be used. This results in a system with an effective 12 hours cut-off time, as observations that arrive with a delay of up to 12 hours are still used.

From a practical point of view, all observations within the window can be used, or only the newly arrived observations. When the impact of distributing observations between the members of the EDA has been fully understood (section 4.2.3), it might be used in this context, distributing observations between iterations for a given sub-window. At the same time, any given sub-window would be processed many times (12 times in the example given here), resulting in many effective outer iterations. More nonlinear effects could be included in the definition of the cost function which brings another potential advantage.

From a computational point of view, the assimilation would become a service that runs quasi-continuously where an up-to-date analysis is always available. Moreover, because of the large overlap, the guess for any given cycle is already very good and just a few inner iterations are needed per cycle. One advantage is that the minimisation does not have to be particularly fast and the code can be optimised for energy consumption rather than time to solution. Another positive consequence is that such a system would not generate peaks in daily computer usage and therefore puts less pressure on the time critical path. ECMWF is already running a project called the Continuous Observation Processing Environment (COPE) that takes initial observation processing out of the critical path by processing observations as soon as they arrive at ECMWF. It is a possibility that the assimilation system could, at some future date, also be run under the COPE system. It could become possible to provide more frequent updates to users, both for product generation and to support their regional modelling activities.

The increase in the frequency of analysis updates for the high resolution DA needs to be matched by a corresponding increase of the frequency of ensemble DA updates to provide consistent background covariance information. The use of observations in an overlapping system might also affect the definition of the background error covariance matrix and could require adjustment in the ensemble DA system. Research is at an early stage and these aspects will be examined in future research.

4.1.3 Saddle-point algorithm

The perfect model assumption made 4D-Var manageable in the past but also made it sequential in time: the state throughout the assimilation window is computed from the initial condition by integrating the model. Because weak constraint 4D-Var relies on a four dimensional control variable, the state is known throughout the assimilation window, or at least at regular intervals (sub-windows) through the window: the cost function can be evaluated in parallel from those states. As a result, weak constraint 4D-Var is more parallel than the standard strong constraint 4D-Var.

In practice, the minimisation algorithm at the heart of variational methods needs preconditioning. Unfortunately, efficient preconditioners for weak constraint 4D-Var, in its model space formulation, are sequential. Another formulation for weak constraint 4D-Var has been developed to preserve parallelism, using a saddle-point optimisation algorithm. Details of the algorithm were presented to the SAC (ECMWF/SAC/40(11)5 2011) and updated more recently (Fisher and Gürol 2017; Fisher, Gratton, Gürol, Trémolet and Vasseur 2016). We only summarise here some important properties of this algorithm:

• The tangent linear and adjoint integrations can run in parallel over the sub-windows.

• Each tangent linear integration can run in parallel with the adjoint integration over the same sub-window.

• The algorithm only uses the covariance matrices (B, Q, R) themselves, not their inverses.

These properties have many implications. Relative to the current implementation of 4D-Var, the degree of parallelism is increased by twice the number of sub-windows (that is already a factor of two in strong constraint 4D-Var). One major hurdle in achieving good scalability in NWP is the fact that when resolution increases, the length of the time step decreases, and the number of sequential time steps increases. Here, the sub-windows can in theory be as short as a time step, with the consequence that available parallelism increases in both the spatial and temporal dimensions as resolution increases: 4D-Var becomes more scalable than the forecast. The difficulty remains that the model error covariance Q is poorly known and shorter sub-windows imply a greater dependency on it. However, because the saddle point algorithm only requires Q and not its inverse, it is possible to make Q go to very small values. This result can also be obtained by writing strong constraint 4D-Var as a constrained optimisation problem and using Lagrange multipliers to impose the model constraint. Weak constraint 4D-Var then converges towards strong constraint 4D-Var but benefiting from additional parallelism. The model error covariance Q can then be made progressively more realistic in the future as science improves. The fact that the inverse of R is not needed also makes this algorithm much more amenable to using correlated observation errors.

Overall, the saddle-point algorithm makes 4D-Var very scalable and has other properties that make it useful. This algorithm is implemented in OOPS where additional preconditioning has already been developed with simple models. It remains a high priority to implement IFS under OOPS to benefit from these advances.

4.1.4   Improved preconditioning

In the current IFS 4D-Var implementation, two levels of preconditioning are used when minimizing the cost function. The first level uses the background error covariance matrix as a preconditioner. The second level relies on properties of the Lanczos algorithm to build a preconditioner as a by-product of each minimisation to precondition the minimisation in the next outer iteration. The idea behind this technique is that information about the shape of the cost function is collected during the minimisation and then used to precondition the minimisation in the following outer loop.

One fact to note is that the definition of the Lanczos preconditioner depends on the trajectory, but in most cases does not depend on the values of the observations. Only their locations and associated errors (through the linearised observation operator H and the observation error covariance matrix R) are required. In theory, the position of most future observations can be known: satellite trajectories are not random, weather stations positions are known, and flight schedules are usually known. By generating this information in the appropriate form, it would be possible to pre-compute the preconditioner as soon as the background is available. Depending on the choice of preconditioner, applying it is exactly or approximately equivalent to repeating the minimisation that lead to its estimation (Trémolet 2008, section 4). This is equivalent to pre-computing the gain and would essentially move the first minimisation out of the time critical path.

In reality, the position of some observations is not easily predicted with precision (e.g. radio occultations, drifting radiosondes) and some observations affected by non-linear on-off processes may have a wrong observation operator (e. g. clear instead of cloudy radiance observations). Communication links can fail, some observations will be removed by quality control, some stations might miss a report, flights might be delayed or cancelled (we all have been stranded in an airport at some point!). However the vectors appearing in the expression for the preconditioner are related to the eigenvectors of the Hessian of the cost function which are large scale. Typically, the leading eigenvectors are related to surface pressure observations over Europe, or aircraft data over North America or a satellite track (Andersson, Fisher, Munro and McNally 1999). As a consequence, they will not change significantly if one station report is missing or a flight is delayed. Pre-computed values should still be valid, and maybe a few additional iterations of the minimisation algorithm with the actual observations could be run to alleviate these differences and finalize the computation of the increment.

We should also note that the EDA members are each solving very similar minimisation problems in parallel. As a result, they all collect information about the shape of the cost function that could be exchanged and used to speed up the convergence of all the members. This technique is well known in the mathematical community as block preconditioning. The main drawback would be that the EDA members become more closely tied together in the minimisation stage so combining the block preconditioning with pre-computing might be useful. On the other hand, if the EDA members are so similar they generate the same information, there is potential for very large efficiency improvements by pre-computing the gain only once for all members. Given the cost of the EDA, it is important to investigate the potential of such algorithms. It should be noted that the block preconditioning technique can also be applied to En-4D-En-Var.

4.1.5   The Linearised and Adjoint Models

In 4D-Var, the minimization of the cost function requires the use of linearized versions of the reference non-linear (NL) forecast model, the so-called tangent-linear (TL) model and its mathematical transpose, the adjoint (AD) model. While the TL model is integrated to describe the time evolution of the analysis increments, the AD model allows the conversion of the gradient of the 4D-Var cost function in observation space and at observation time into a gradient of the cost function with respect to the model state variables (temperature, moisture, wind and surface pressure) at the beginning of the assimilation window.

Because the 4D-Var minimization can only be solved iteratively (using of the order of 30 iterations, in the current ECMWF 4D-Var), to make 4D-Var affordable the linearized model needs to be computationally efficient. Since 4D-Var relies on the central assumption that all processes described during the minimization should be (quasi) linear, not all physical processes can be included in the linearized model. For instance, particular issues may arise when dealing with processes that involve highly nonlinear behaviours (e.g. transitions between stable and unstable boundary layer near the surface) or thresholds and switches (e.g. saturation). Fulfilling the requirements of affordability and linearity would advocate the development of a linearized physics package that is simplified relative to the reference NL model. However, in order to ensure the best quality of 4D-Var analyses, it is also necessary that the linearized schemes: 1) describe the atmospheric behaviour in a way as realistic as possible; and 2) match their NL counterparts as closely as possible (the so-called tangent-linear approximation). A delicate balance must be achieved between these two sets of opposite requirements.

The initial implementation of the linearized model at ECMWF (Buizza 1993) was limited to the dynamic core of the NL model plus a highly simplified vertical diffusion scheme. This initial formulation has been progressively developed and extended to represent all the significant physical processes described by the NL model (Mahfouf 1999; Tompkins and Janisková 2004; Lopez and Moreau 2005; Orr, Bechtold, Scinocca, Ern, and Janisková 2010, Janisková and Lopez 2013). The current linearized model has thus reached a level of sophistication which is comparable to that of the reference NL model and which is not found in other 4D-Var implementations. This can be shown to have a clear positive impact on 4D-Var accuracy (Figure 1; see caption for explanation).

The level of detail of the linearized physics package has also a direct impact on which types of observations can be successfully assimilated in the 4D-Var system. For instance, the operational assimilation of satellite microwave brightness temperatures in cloudy or rainy situations and of ground-based radar precipitation composites (Lopez 2011) can only work because moist physical processes are part of the linearized model (otherwise the gradient of the cost function associated to these observations would be zero). Besides, recent developments have been aiming at the inclusion of reflectivity and backscatter computations in the linearized physics package to allow the direct 4D-Var assimilation of space-borne cloud-radar and lidar data from the future EarthCare mission (Janisková 2015). Furthermore, the use in 4D-Var of lightning observations from the next generation of geostationary satellites (GOES-R, FY-4A and MTG) and from ground-based networks will only become possible if a lightning parameterization is added to the linearized physics package (Lopez 2016). Eventually, only the high level of sophistication achieved by the ECMWF linearized model can help maximize the 4D-Var exploitation of the wide range of observation types that are currently available on the global scale.

Of course, the careful maintenance of such a complex linearized system is paramount in order to constantly optimize the performance of 4D-Var. Every time the TL model is changed, any instability resulting from the spurious growth of any perturbations must be identified and eliminated. Furthermore, every time a new version of the reference NL model is developed (i.e. typically for each new model cycle), with frequent changes in dynamical and physical computations, the validity of the tangent-linear approximation needs to be re-evaluated in special offline tests and the scores of 4D-Var experiments need to be monitored. If some degradation is found in the TL approximation or in the 4D-Var scores, the TL, AD or even sometimes reference NL model codes must be modified. Solving such issues can be sometimes demanding, due to the growing complexity of the IFS, especially when operational deadlines are to be satisfied. Despite that, one should note that in practice the workforce dedicated to the linearized physics at ECMWF is limited to 1 person-year.
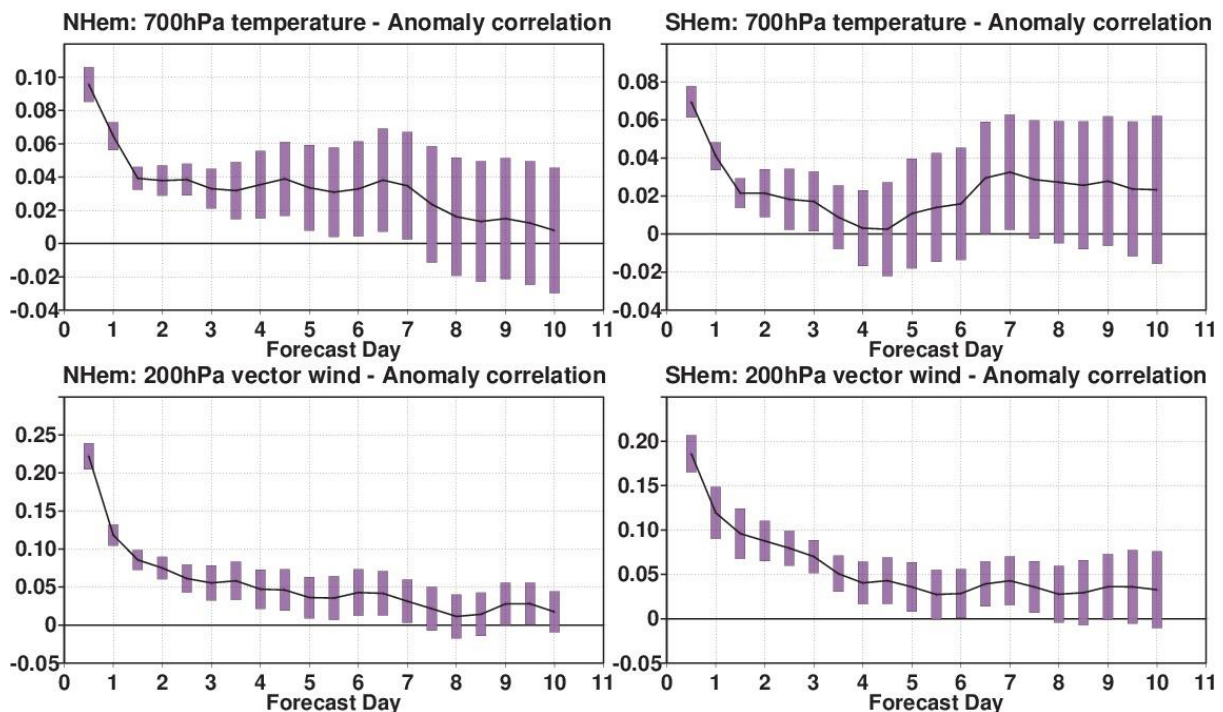

Figure 1: Relative change in forecast anomaly correlation resulting from the inclusion of linearized physics in the 4D-Var minimizations, from two T511 L91 4D-Var experiments over the period July-September 2011. Positive values indicate better skill when the linearized physics is included in 4D-Var, vertical bars indicate 95% confidence levels. Displayed parameters are 700-hPa temperature (top row) and 200-hPa wind vector (bottom row), over the northern and southern hemispheres (left and right columns, respectively).

In the future, continuous efforts will be devoted to maintain the good performance of the linearized model (both physics and dynamics), in particular every time horizontal resolution is increased not only in the 4D-Var minimizations but also in the 4D-Var trajectories. Recent experience has shown that special attention has to be paid to the linearized version of the dynamical core to remove instabilities that can develop when 4D-Var minimizations are run at horizontal resolutions of 25 km and higher. With regards to the linearized physics, when 4D-Var trajectories are run at kilometre resolution, the

impact of the different treatments of convection (resolved in the trajectories versus parameterized in the lower-resolution minimizations) on the validity of the tangent-linear assumption will have to be assessed. In the medium term, one will also have to investigate whether the resolution of 4D-Var minimizations can be increased to the kilometre-scale. However, until then and as emphasized earlier, the cost of maintaining the linearized physics package is expected to remain reasonable, even though increased resources might be required to cope with the increasing complexity of the IFS or if more processes are to be represented (e.g. coupling with the surface). Overall, linearized model developments will be pursued in view of the clear impact they have on the 4D-Var performance.

It should also be noted that the ECMWF linearized model is also used in other applications beyond 4D-Var: singular vector computations for the generation of initial perturbations in ensemble forecasts and in the computation of the sensitivity of forecasts to observations (Janisková and Cardinali, 2017).

4.2   EDA

One limitation of the current hybrid 4D-Var based on the EDA error cycling resides in the significantly lower quality of the EDA members with respect to the high resolution 4D-Var analysis. This gap in accuracy is mainly due to two factors: the reduced resolution (both outer and inner loops) at which the EDA members are run to make computational costs affordable; and the application of explicit perturbations to the model and the observations. There are ways however to mitigate these problems in the EDA: we discuss two possible strategies below. In addition we will discuss multi-resolution ensembles as further option to reduce cost.

4.2.1   Asymmetric EDA

It is common for various implementations of ensemble Kalman filters to separate the analysis into the calculation of an ensemble mean analysis and of a set of ensemble mean analysis perturbations. Lorenc, Jardak, Payne, Bowler and Wlasak (2016) applied a similar approach to an ensemble of variational analyses. The authors separate the analysis into an ensemble mean analysis calculated using 4D-En-Var and the deviations from the ensemble mean analyses calculated using linearized equations and 3D-Var. The mean-perturbation method was demonstrated not to alter significantly the RMSE of the ensemble mean and RMS ensemble spread for experiments presented in Bowler, Clayton, Jardak, Lee, Lorenc, Piccolo, Pring, Wlasak, Barker, Inverarity and Swinbank (2016a), while the analysis cost was reduced by a factor of 3. Buehner, McTaggart-Cowan and Heilliette (2016a, b) also adopted a similar approach and proposed further simplifications to the perturbation analyses by reducing the amount of assimilated observations and using a climatological covariance B (effectively running 3D-Var analyses).

Separation of the analysis into the ensemble mean and ensemble mean deviations amounts to making an implicit assumption about the Gaussianity of the posterior pdf and linearity of the operators. Here, a similar approach to the one proposed by Lorenc, Jardak, Payne, Bowler and Wlasak (2016) is explored in the EDA. The experiments have an asymmetric outer loop configuration for the EDA control and perturbed members and an intrinsic re-centring step. Such configuration allows to reduce the EDA's cost and retain its attractive property of representing the evolution of the high-resolution 4D-Var assimilation errors.

In the operational configuration of the EDA, the control and the perturbed members are run with two outer loops. In the configuration proposed here the aim is to calculate a more accurate analysis for the

control member, while the accuracy of the perturbed ensemble member analyses is reduced. In this test setup the control member is run with 3 outer loops and the perturbed members are run with only one outer loop. To initialise the first guess forecasts, the perturbed members are then re-centred on the control member. A schematic of the re-centring step is shown in Figure 2.

Results are shown for two EDA experiments, a reference (ref) and an asymmetric EDA experiment with re-centring (rec). Both are run at TL399 resolution with 25 members. The inner loop resolutions for the reference experiment are TL95 and TL159. For the asymmetric experiment with re-centring the perturbed members are run with one outer loop (inner loop resolution TL159) and the control is run with three outer loops (inner loop resolutions TL159, TL159 and TL255). The experiments were run from 1 June 2015 to 13 July 2015. The scores of the forecasts started from the control member improve markedly for the experiment with re-centring (Figure 3). This shows how large is the impact of a more accurate control member configuration on the forecast skill.
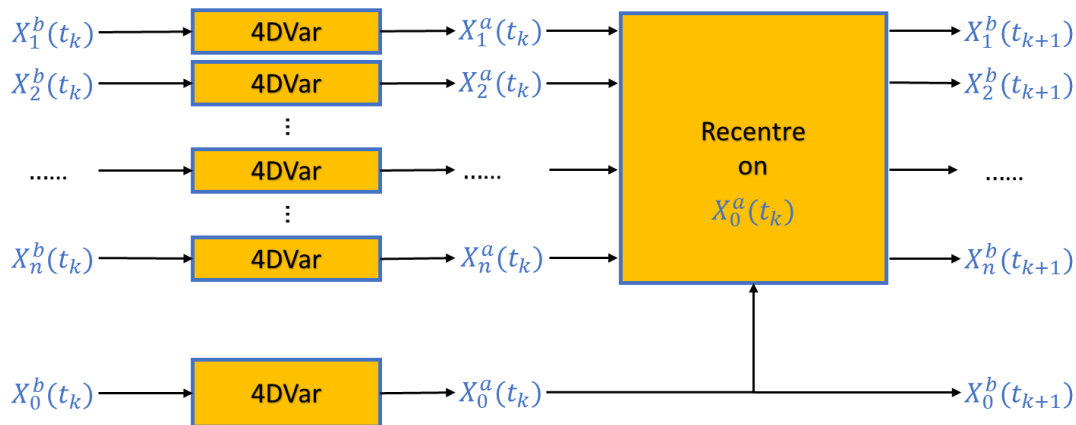

Figure 2: Schematic showing the re-centring step. $X_0^a(t_k)$ denotes the control member analysis.

In terms of cost, a first estimate is that this setup can be approximately 40% cheaper than a reference configuration (depending on the reference EDA setup; note that in the operational TCo639 EDA both inner loops are run at TL191 resolution). This cost saving could allow an increase in the number of EDA members and allow to put more emphasis on the flow dependent covariance estimates.

It is worth noting that the asymmetric EDA configuration with re-centring is a fully 4D-Var based system. It is also in line with ECMWF strategy on model uncertainty (e.g. seamless model error representation: ECMWF/SAC/45(16)8 2016Leutbecher et al., 2016) and diagnostics (e.g. reliability budget, Rodwell, Lang, Ingleby, Bormann, Hólm, Rabier, Richardson and Yamaguchi 2015). The system design is flexible allowing to easily change configuration, e.g. in case more computing resources become available.

In general a re-centring step is not desirable from a theoretical point of view because it is based on a linearity assumption. An attractive property of the EDA is that it can handle weakly non-linear situations at analysis time and this can lead to improved initial conditions for ensemble forecasts (Lang, Bonavita and Leutbecher 2015). However, for this to be possible a high quality EDA would be needed in real-

time to initialise the ensemble (currently started from the high resolution analysis with perturbations from 6h-EDA forecasts). Lang, Bonavita and Leutbecher (2015) show that, as long there is a clear difference in terms of analysis quality, the gain from re-centring on a better quality analysis can still outweigh these disadvantages. In times of constrained computer resources an asymmetric EDA configuration with re-centring could be a viable option allowing to increase the number of EDA members and thus to improve the flow dependent covariance estimates computed from the EDA first guess forecasts and the ENS initial conditions.

Future plans include testing the impact of the variance and correlation estimates from the two experiments on higher resolution assimilation experiments, similar to the operational configuration of the ECMWF DA system. Furthermore, it is planned to explore the impact of adding more ensemble member in the EDA experiment with re-centring and to test the impact of the perturbations computed from the experiments on ensemble forecasts. In addition, it is important to test ways to further reduce the cost of the system, e.g. use the first minimization of the control member to precondition the perturbed members. If the preconditioning of the perturbed members turns out to be successful it might also be possible in the future to run an up-to-date EDA in time for initialising the ensemble forecast, possibly with a control running in quasi-continuous mode and the members with a pre-computed gain (see Section 4.1).
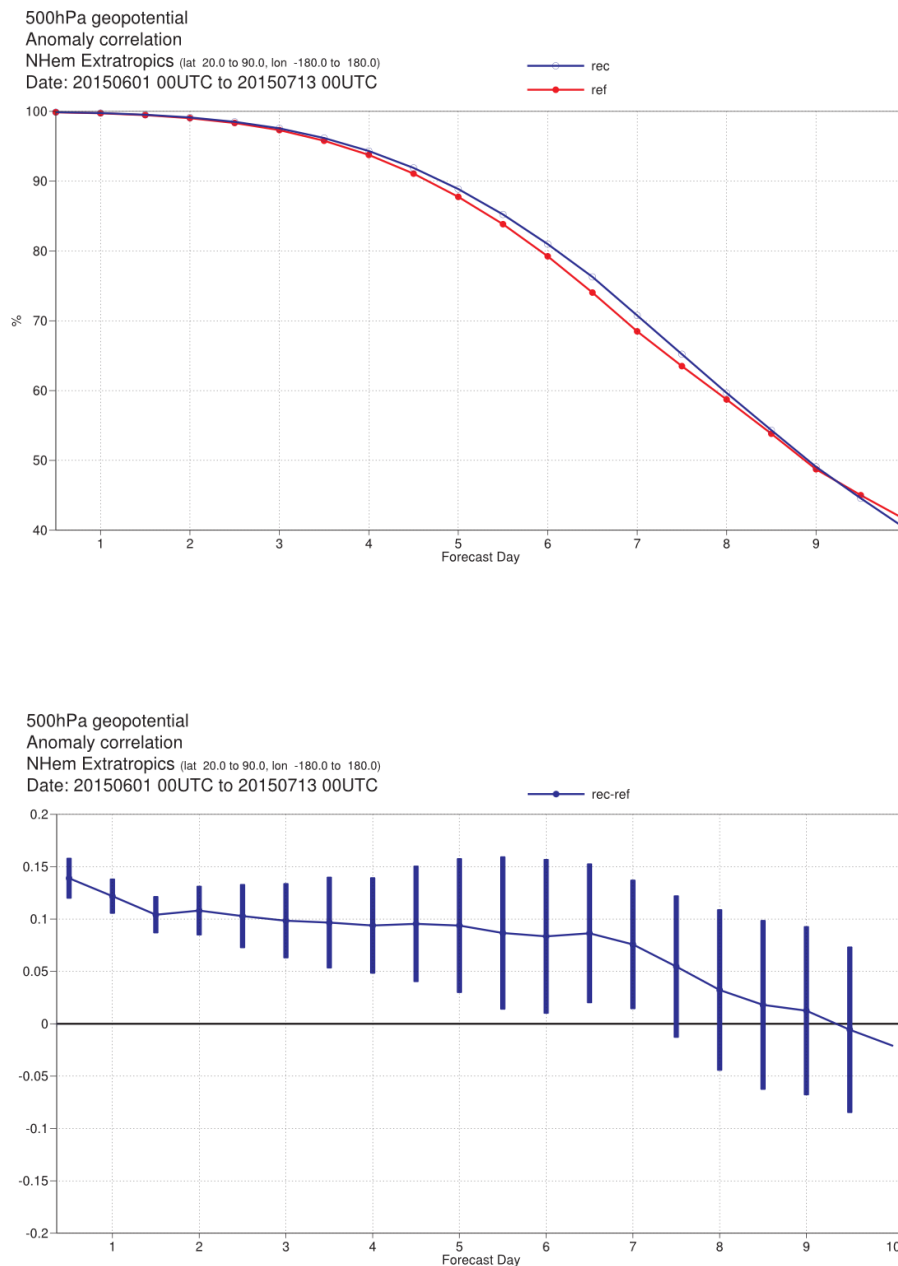
Figure 3: Top panel: Geopotential forecast anomaly correlation at 500hPa for the control member of experiment ref (red) and rec (blue); Bottom panel: Normalised differences of forecast anomaly correlation of the control members. Scores are averaged over the period 20150601 00UTC to 20150629 00 UTC, forecasts started every 24h.

4.2.2   Multi resolution EDA

One way to reduce the computational cost of higher resolution ensembles is to run multi-resolution ensembles. This has been tested by e. g. Rainwater and Hunt (2013) for an LETKF and Gregory, Cotter and Reich (2016) for a particle filter. The latter uses the multilevel Monte Carlo method of Giles (2008)

which has been developed into a framework for optimizing the resolution and number of members in a series of sub-ensembles to obtain best accuracy for a given computational cost. For the EDA a series of sub-ensembles at different resolutions can be used to more effectively sample the covariance matrix. In order to effectively make use of multi-resolution ensembles for sampling the covariances, it is necessary to average the highest resolution members over a relatively large area in order to increase sample size for the highest wavebands of the wavelet B. As an example, the highest waveband used operationally at ECMWF is at TL399/50km, with the spatial samples averaged to a 250km resolution for B_{TL399}, which allows almost complete flow-dependency at this scale with only 7% contribution from the climatological covariance matrix.

As pointed out by Berre, Pannekoucke, Desroziers, Stefanescu, Chapnik and Raynaud (2007), to effectively increase the number of correlation matrix samples, the averaging length scale should be larger than the sample noise correlation length scales while at the same time the correlations should vary slowly over the averaging area. Recent work on combining a diagonal wavelet formulation with local grid deformations that make the correlations more isotropic show that this is in principle possible even on the convective scale (Michel 2013a, b; Legrand and Michel 2014; Pannekoucke, Emil and Thual 2014). Adding isotropizing grid deformations (local elongation and rotation) to a diagonal wavelet formulation of B brings it closer to a full anisotropic grid point correlation while simultaneously enabling spatial averaging over larger areas to increase sample size. This is because the correlations are calculated on the transformed grid where they vary slowly. In this way the flow-dependency of the smallest scale correlations are moved from the correlation matrix to the grid-transformation.

In addition a multi-resolution ensemble provides flexibility to adjust available resources for changes in balance between memory and CPU constraints. A further aspect of a mixed resolution ensemble is that it could provide an explicit sampling of the resolution-dependent part of model uncertainty. How to use multi-resolution ensembles to initialize ensemble forecasts is an open question, ranging from running a multi-resolution ensemble to methods combining different resolution EDA members for the initial conditions.

4.2.3   Distributed observations

The EDA can be considered a variational implementation of the perturbed observation EnKF, where observations are perturbed according to their assumed observation error covariances. The observation error covariances are often poorly known, especially in terms of their correlation structure. As the total volume of available observations is constantly growing but only a few percent of it is currently used, an alternative approach to ensemble data assimilation in the EDA (or perturbed observations EnKF) context is to use different subsets of all the available observations in different members. This would have the advantage of eliminating a source of sampling error in the system.

Compared with a standard EDA/perturbed observation EnKF, we replace random observation perturbations drawn from the assumed observation error distribution by a random sampling from the available observation sample. As a first step to explore this idea, four analysis experiments were run for 50 days with each using distinct subsets of infrared and microwave satellite, scatterometer and atmospheric motion vector wind observations, with all other observations the same as in the control, and without any perturbation to the forecast. The forecast scores from each of the four analyses are not distinguishable from those of the control analysis at the same resolution, as seen in Figure 4 (blue line).

We quantify the contribution of the perturbations to the analysis quality by running deterministic forecasts from five different experiments: standard observations (reference), selected observations (blue line), selected observations and perturbed model (green line), perturbed standard observations (red line), and perturbed standard observations and perturbed model (like current EDA, black line). In contrast to selecting different subsets of observations, perturbing the observations or the model is seen to degrade scores significantly.
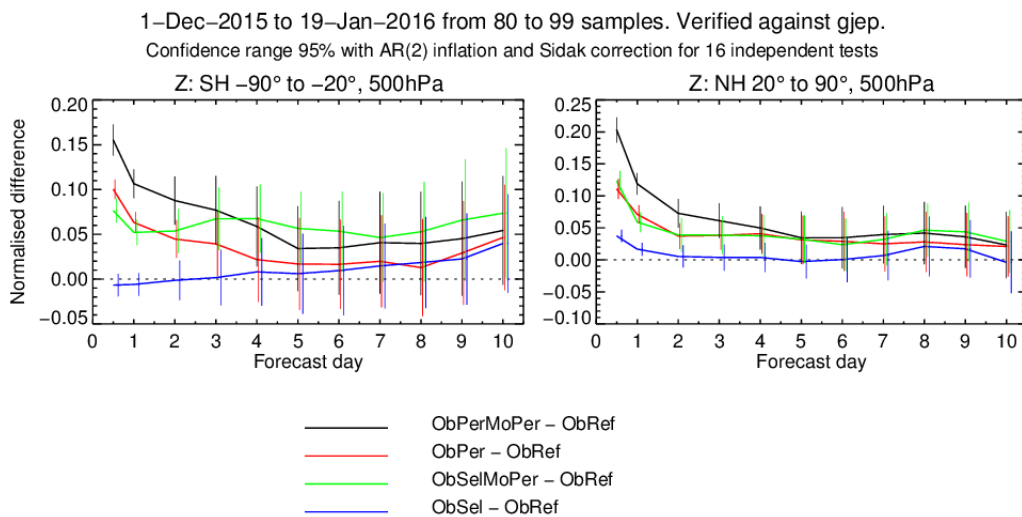

Figure 4: Contribution of EDA observation and model perturbations to 500hPa RMSE for Southern (left) and Northern (right) Hemisphere. Black: perturbed standard observations and model. Red: perturbed standard observations. Green: selected observations, perturbed model. Blue: selected observations. Zero line: standard unperturbed observations and model. All assimilation experiments are at TCo639 resolution with TL255-319-399 inner loops and all forecasts are run from 06/18 and verified against the reference analysis.

The spread is qualitatively similar to the EDA spread, with similar amplitude in the Tropics but less amplitude in the extra-Tropics (Figure 5). Further research in this area will evaluate different possibilities for distributing the observations between the members (see He and Garcia 2009, for a review of an analogous problem in ensemble machine learning). One method is to maintain the same observation density as in the control analysis for all observation sets, thus gradually going from unique sub-sets for very dense observations (e.g., hyper-spectral sounders, where we can sub-sample in both physical and radiance space), then randomly overlapping sub-sets for medium density observations (e.g., scatterometers) and finally using identical sets in all members for sparse observations (e.g., drifting buoys) and anchoring observation sets for bias-correction (e.g., radiosondes, limb sounders).

The use of different observations in different members is related to the distributed Kalman filter (see for example Ajgl, Šimandl and Dunik 2009; Mahmoud and Khalid 2013), where one central analysis of all observations is replaced by several distributed analyses. In the distributed Kalman filter several methods have been developed how to best combine the distributed analyses (e. g. Shin, Lee and Choi 2006). These methods can be applied in our applications to combine the analyses from the different members with un-perturbed observations, resulting in a merged analysis that contains all available observations

and possibly has lower error than the control. This is different from the standard EDA where the ensemble mean analysis is of the same quality as the control for linear error evolution.
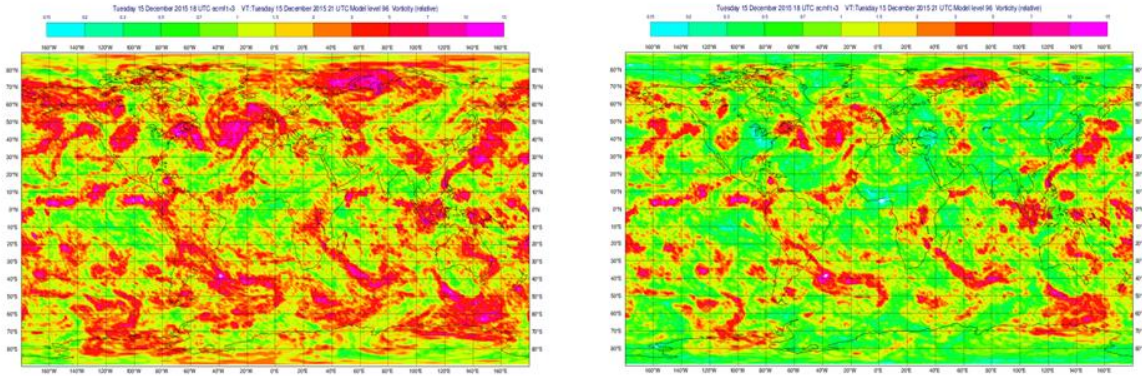

Figure 5: Vorticity spread at model level 96 (ca. 500 hPa, unit 10-5 s-1) from four non-perturbed observation analyses with different observation selection and no model error perturbations in the forecast (right) versus spread calculated from four EDA members (left). Each analysis uses distinct subsets of satellite radiances (all-sky and others), scatterometer and atmospheric motion vector winds, with all other observations the same as in the control.

4.3 Multiple data assimilation approaches: Hybrid Gain

An EnKF-based error cycling system would ideally be able to retain the computational efficiency of the square root version of the EnKF for computing the analysis perturbations and at the same time avoid the EnKF limitations with regards to sampling errors and localization. This has motivated research into the Hybrid Gain Ensemble Data Assimilation (HG EnDA) system, based on the initial concept proposed by Penny (2014). The schematic of the HG EnDA is shown in Figure 6.
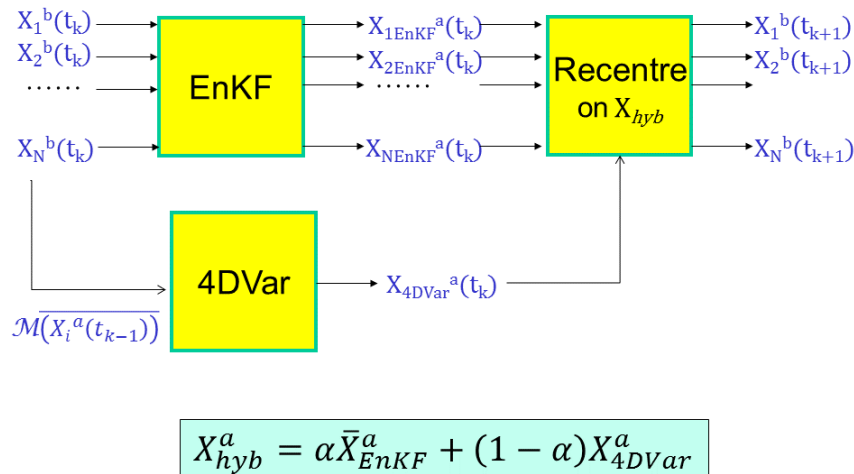

$$X_{hyb}^a = \alpha \bar{X}_{EnKF}^a + (1 - \alpha)X_{4DVar}^a$$

Figure 6: Schematic of the Hybrid-Gain EnDA system.

From this diagram we can see that in addition to a standard EnKF analysis update, an incremental 4D-Var analysis is also performed. The 4D-Var analysis is tightly coupled to the EnKF analysis: it uses the short range forecast from the previous ensemble mean analysis as the background state valid at the beginning of the assimilation window and as a first-guess linearization trajectory. The EnKF and 4D-Var analysis increments valid in the middle of the assimilation window (i.e., 00, 06, 12, 18UTC in the 6 hour cycling set-up used in the experiments) are then linearly combined to produce a control analysis, around which the analysis ensemble is centred. Note that as the background state of the EnKF and 4D-Var is the same, the re-centring step is effectively a blending of the two analysis increments. Differently from a standard hybrid B data assimilation system where a linear combination of an ensemble-based B and a static B is used in the analysis update, in the HG EnDA a linear combination of a full ensemble-based and climatological Kalman Gain matrices is computed. The weight given to each of the contributing analysis increments (denoted as α) is a tuning parameter, reflecting the expected accuracy of the EnKF and 4D-Var analyses. A value of α=0.5 has been used in the experiments reported in this work, i.e. the 4D-Var and EnKF control analysis are given equal weight. This is thought to be not far from the optimal value for a constant α, as later experimentation with α=0.75 and α=0.25 produced slightly inferior results. In general the optimal weight should however vary with the error covariances of the two analyses (Shin, Lee and Choi 2006).

Results reported in Bonavita, Hamrud and Isaksen (2015) showed that the HG EnDA control analyses and forecasts were significantly more accurate than either their EDA or EnKF counterparts at TL399 resolution. More recent experimentation at the current operational resolution of the EDA (TCo639) confirms that forecasts started from the HG EnDA control analysis can give similar scores to those started from the current operational high resolution 4D-Var analysis (Cycle 41r2; TCo1279), Figure 7. This result suggests that there is still room for improvement in the high resolution analysis from a tighter coupling with the ensemble DA component and a more direct use of ensemble information in the analysis algorithm.

In the experiments reported here, the 4D-Var component of the HG EnDA has been run using climatological estimates of the background error covariances and at relatively low resolutions for the inner loop minimizations (TL95/TL159 in the experiments described in Bonavita, Hamrud and Isaksen (2015); TL95/TL159/TL255 in the experiment shown in Figure 7). With this setup the 4D-Var analysis produces relatively large scale, smooth analysis increments (Bonavita, Hamrud and Isaksen 2015, Figures 13, 15) and thus acts to: a) to reduce the effect of sampling noise and localization in the EnKF, and b) to efficiently introduce climatological information in the background error covariance estimates, which are completely flow-dependent in the EnKF component.

It is worth noting that the performance of the HG EnDA has been substantially improved by the recent addition of cloud and precipitation affected radiances (all-sky radiances in the following; Bauer, Geer, Lopez and Salmond 2010). These observations were withdrawn in previous experimentation due to concerns over the behaviour of the EnKF when faced with observations with a strongly non-linear dependence on the background state. Recent research has however demonstrated that the EnKF is able to make effective use of these data, provided: a) the non-linearity of the problem is explicitly accounted for in the observation error model, and b) care is taken in the localization of the impact of all-sky observations in the analysis to reflect their more local signature with respect to, e.g., temperature sounding channels. The sensitivity of the EnKF to adding the all-sky observations to the rest of the

observing system has in fact been found comparable to that seen in the 4D-Var context (Geer, Baordo, Bormann and English 2014).
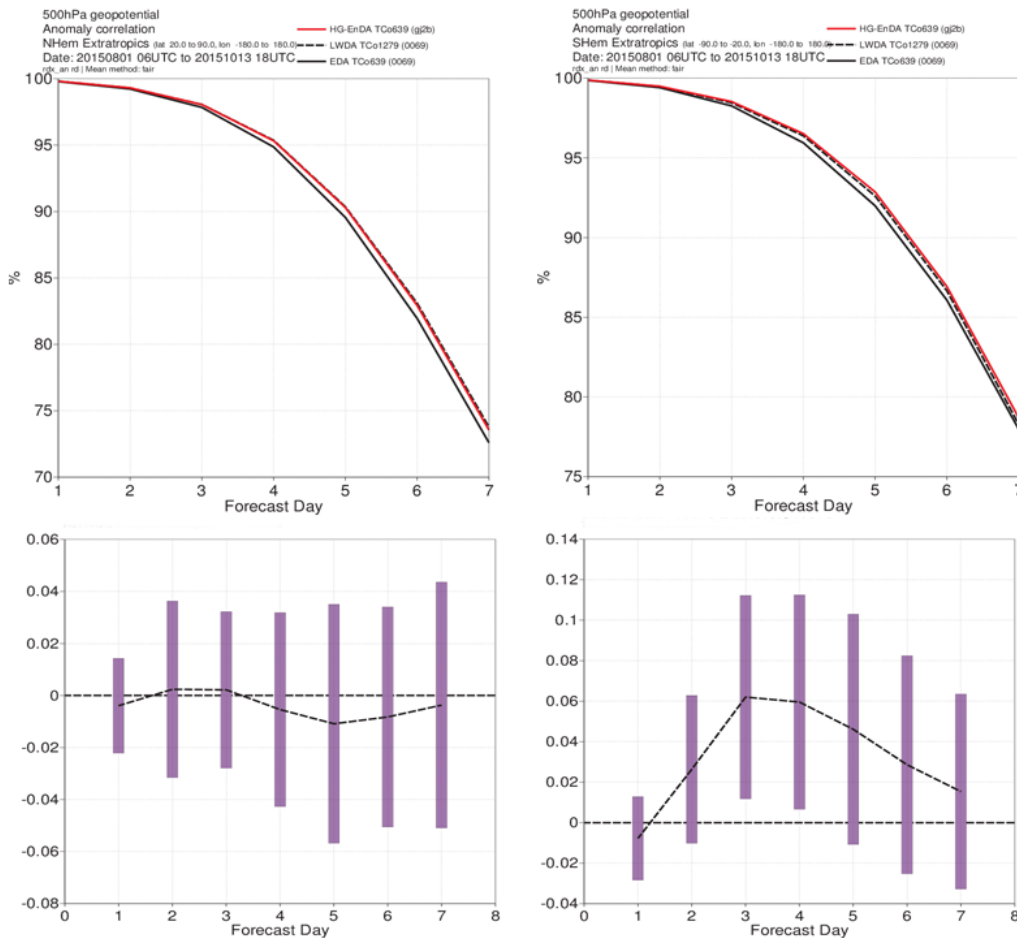

Figure 7: Top panel: Geopotential forecast anomaly correlation in the northern hemisphere (left panel) and southern hemisphere (right panel) at 500 hPa of forecasts started from: the control analysis of a TCo639 100 member HG EnDA with allsky radiances (continuous red line); the control analysis of the 41R2 esuite TCo639 EDA (continuous black line); the 41R2 esuite high resolution TCo1279 analysis (dash black line). Bottom panel: normalised difference of anomaly correlation forecast scores from the HG EnDA and the high resolution forecasts (positive values indicate better skill of the HG EnDA). Scores are computed with respect to ECMWF operational analysis and are averaged over the 2015-08-01 to 2015-10-13 period.

The statistical characteristics of the ensemble perturbations produced by the HG EnDA are very similar to those produced by the EDA, both in terms of spatial distribution and spectral signature. This implies that the background errors and covariances estimated by the two systems are also very similar.

A sensitivity experiment has been conducted to test the impact of using background errors and error covariances from a HG EnDA system in a high resolution (TCo1279) data assimilation cycle run in a configuration close to that currently used in operations (IFS Cycle 41r2). Results compared to the pre-operational implementation of Cycle 41r2 appear generally neutral (not shown).

5   Coupled Data Assimilation

5.1   Introduction and current status

The drive towards a coupled data assimilation and forecasting system for the full Earth system, comprising atmospheric composition, land surface, ocean and sea ice components (ECMWF Strategy 2016-2025) is an exciting development in current NWP. It raises a number of new challenges for data assimilation as each of these components will require to be initialised in a physically consistent manner with all the others.

There are many potential benefits to coupled data assimilation for both NWP and reanalysis applications. For example, it should reduce surface initialization shocks in the forecasts. Another interesting prospect would be the possibility of producing consistent SST and sea-ice analyses for ocean and atmosphere. A coupled ocean analysis should produce more consistent SST analyses than an external one as currently available by virtue of combining the oceanic in-situ observations, satellite measurements, atmospheric winds and ocean mixed layer information. This will rely on both the development in NEMOVAR of the assimilation capability and the timely provision of bias corrected level 2 SST products. Retrievals of quantities such as sea ice concentration should also benefit. Figure 8 shows an example of improved use of observations in a coupled data assimilation system.

In the 4D-Var framework a range of coupling strategies ranging from the weakest (separate assimilation for each component but with forecasts using a coupled model) to the strongest (a single cost function and a control vector that includes increments for all models) can be considered. Between these extremes, a first approach has already been demonstrated in a coupled Ocean-Atmosphere 4D-Var context with the coupled reanalysis project (CERA, Laloyaux, Balmaseda, Dee, Mogensen and Janssen 2016) where the 4D-Var outer loop is run with the coupled Ocean-Atmosphere model (so that the coupled models can adjust to each other during each outer-loop forecast) but the minimization for each sub-system is independent. Production of the CERA-SAT coupled reanalysis covering the satellite era will be based on this system. An extension of the reanalysis production to the near real time in the future is envisaged although access to oceanic observations within operational data cut-off times will be a challenge as it could be for a future coupled operational system.

The land surface data assimilation is another example where data assimilation coupling is also progressing. Recent developments of the simplified extended Kalman filter (SEKF) make use of the EDA perturbations to estimate the surface analysis Jacobians, as shown in Figure 9. The EDA Jacobians (bottom) are noisier than the finite difference Jacobians (top). However, there is a good agreement in terms of pattern and sign and preliminary experiments indicate promising results. Further developments are ongoing to enable an operational use of the hybrid EKF-EDA soil moisture analysis. This approach leads to a substantial reduction in computing cost, compared to the finite difference approach used in operations (de Rosnay, Drusch, Vasiljevic, Balsamo, Albergel, and Isaksen 2013), and it opens the possibility for coupling at the outer loop level and use of flow dependent background errors.

Three main aspects should be considered when moving forward in the domain of coupled data assimilation: the use of observations that depend on more than one system, the propagation of coupled increments and the definition and use of cross-covariances. Each can be developed and used independently but ultimately all aspects should be used together for an optimal coupled DA.
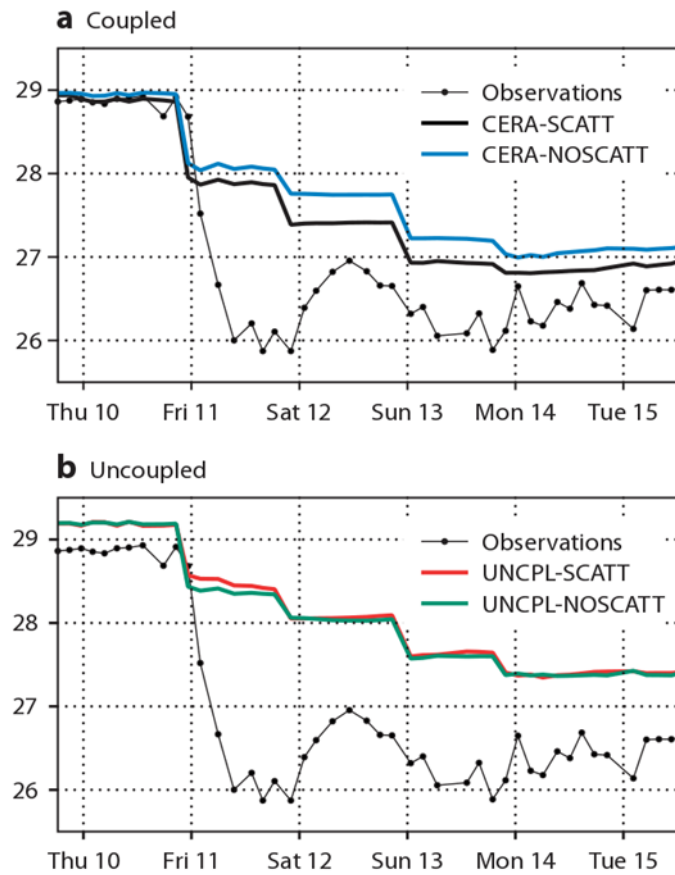
Figure 8: Time series of ocean temperature observations at a depth of 40 metres measured by an Argo float during the passage of cyclone Phailin with (a) the temperature analyses produced by the CERA system with scatterometer assimilation (CERA-SCATT) and without scatterometer assimilation (CERA-NOSCATT); and (b) the temperature analyses produced by the uncoupled system with scatterometer assimilation (UNCPL-SCATT) and without scatterometer assimilation (UNCPL-NOSCATT). In the coupled assimilation system, scatterometer data (CERA-SCATT) led to a consistent improvement in the temperature estimate, by up to 0.4°C, compared to the analysis without scatterometer data (CERA-NOSCATT), while the impact is neutral in the uncoupled system.
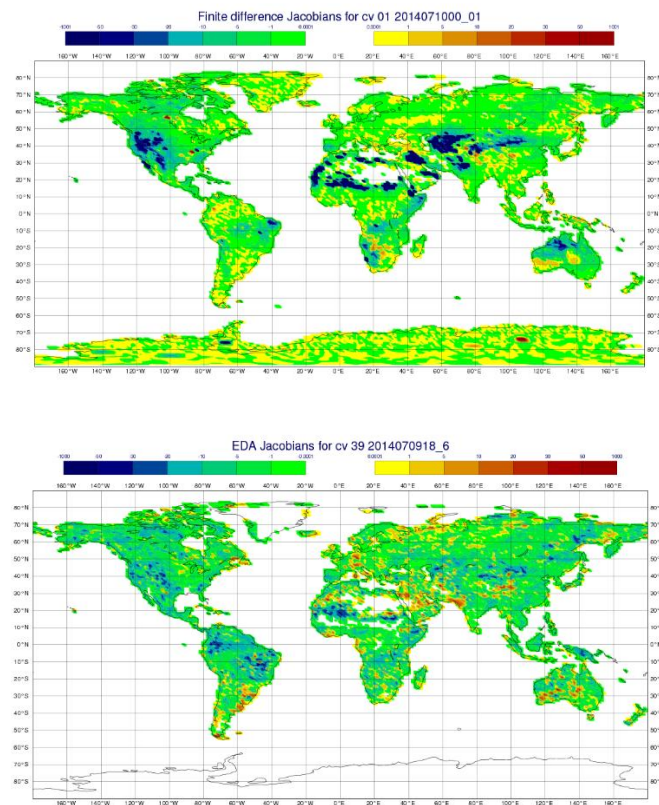
Figure 9: Jacobians matrix elements, in $K/(m^3.m^{-3})$, of the ECMWF top layer soil moisture analysis for two-metre temperature observations computed (a) in finite differences using perturbed trajectories, (b) using the EDA spread to compute top soil moisture variance and co-variance with two-metre temperature. Conditions of low soil moisture variance and frozen conditions are filtered out to compute the EDA Jacobians. The example is shown for 10 July 2014, using T399 experiments.

5.2   Research perspectives: observations

One often mentioned advantage of a coupled data assimilation system is that an observation in one sub-system can inform about the state of another sub-system through the coupling, and an example of this has been given in the previous section. However, there are also observations that depend on more than one sub-system. Most radiance data, when considering observations sensitive to the lower troposphere, depend on the atmospheric state and on the state of the surface below it. Currently those observations are either not used, in the case that the state of the surface is poorly known, or the data assimilation system tries to fit them by adjusting only one of the states which might lead to spurious solutions. An observation operator that depends on both inputs would find its place in a coupled data assimilation system. It would help to update both the atmospheric and surface states consistently, even if the tangent linear and adjoint models are not coupled and the background error covariance matrix lacks cross-correlations. Some effort has been undertaken to develop such models in the past but the use of physical models has been limited by the inability of NWP systems to provide the fields required by the radiative transfer. This has led to empirical approaches, for example retrieving a surface emissivity from the observations, without any reference to whether this emissivity is physically sensible for the surface in question.

For snow, microwave observations are sensitive to layers in the snow and ECMWF is moving towards multi-level snow models. The case to develop observation operators that can use the fields generated by these new model developments is much stronger now than it has been in the past, and this case will only get stronger in the future. Therefore it is planned to develop new observation operators for snow and other surfaces.

Another example arises from a recent collaboration between Météo-France and ECMWF on use of output from the wave model to assist atmospheric DA. The dissipative wave energy from the wave model was used to calculate white cap foam fraction of breaking waves (Meunier, English and Janssen 2014). Microwave ocean surface emissivity is very sensitive to ocean foam. In RTTOV the foam coverage is estimated from the instantaneous 10m wind speed, but this merely reflects that there is a correlation between strong winds and breaking waves. Using the wave model output directly was an important new advance. The foam coverage estimate can then be used in RTTOV to calculate microwave ocean surface emissivity. Results already show it performs at least as well as the wind speed based approach.

A more mature example is ozone, where ECMWF creates an ozone analysis to support hyperspectral radiance assimilation. In the future this could be extended to other atmospheric constituents, such as aerosol, as the coupled DA and model framework advances.

To support this it will be an important and developing research direction to build observation operators across many such areas (ocean, waves, snow, sea ice, land, atmospheric composition). This will allow us to take full advantage of improved models, and to support the strategic goal of coupled data assimilation.

5.3   Research perspectives: algorithms

In the future, data assimilation coupling in the variational framework will require solving a global problem with a control variable which extends to all the sub-systems. Using coupled tangent linear and adjoint models, 4D-Var can generate increments in one sub-system from an observation in another one, even without cross-covariances in the specified background error covariance matrix. On the other hand, using independent tangent linear and adjoint models with a background error covariance matrix that accounts for cross-covariances would also generate increments in one sub-system from an observation in another one. Using both in principle gives the best analysis.

Another possibility is to define a cost function that is the sum of the 4D-Var cost function for each system (with a background and observation term) and an additional coupling term. The two tangent linear and adjoint models run independently and it is the assimilation algorithm that imposes the coupling during the iterative minimisation process. This last approach accounts for the coupling in the data assimilation, including the fact that this coupling might be imperfect and uses the data assimilation process to estimate the coupling error. Because the coupling is imposed as a weak constraint, it might be easier to compensate for biases that affect one system relative to the other. This is very similar in concept to weak constraint 4D-Var which couple the same model over two adjacent time domains and where the model error covariance matrix acts as a coupling term. However, defining such a coupling term for a cost function coupling different models is new and would require a significant research effort. Writing the tangent linear and adjoint of the coupling processes also requires resources but with a more

predictable outcome. While preliminary research activities are starting in this area (Pellerej, Vidard and Lemarié 2016; Smith, Fowler and Lawless 2015) is not clear at this stage which method will be the most effective and practical.

To complement the propagation of the analysis increment throughout the assimilation window, a variety of coupled background error covariance specifications are possible, ranging from a block-diagonal matrix that ignores inter-model correlations, to more complex B matrix models that include cross correlations between models. Coupled background forecasts can be run from ensembles of analyses even if those analyses are not coupled. This is done for example in the cycling of the land-atmosphere operational system and in the CERA system for the ocean-atmosphere coupling. From such an ensemble of forecasts, it is possible to estimate cross-correlations, either by fitting a model to the sample or by using an augmented (α) control variable approach. This will form a good base to start evaluating the benefit of cross-correlations in the background error covariance matrix even if the tangent linear and adjoint models are not yet coupled. Then, it will be important to evaluate the effect of running the ensemble data assimilation component in coupled mode in order to estimate the additional impact of cycling the errors in a fully coupled system.

The key factor that makes a variational data assimilation system coupled is the use of a coupled state as input for the computation of the various terms of the cost function. Ideally, this coupled state is evolved by a coupled model and inside the minimisation by coupled tangent linear and adjoint models. However, nothing prevents the use of two increments that evolved independently, for example by two uncoupled models, to form a global increment to be used as input for the Jo and/or Jc computations. This can be taken further: in the case of an Ocean-Atmosphere system, the ocean increment can be kept constant for the length of the assimilation window and used for all time steps with an evolving atmospheric state. This results in an atmospheric 4D-Var coupled with an ocean 3D-Var, at inner loop level, with potential for use of cross-correlated background error covariances or coupled observation operators. Extending this approach, the increment for one sub-system could also come from an En-Var if tangent linear and adjoint models are not available or difficult to obtain for one of the sub-systems such as a sea-ice or chemistry model.

Yet another form of coupling could be implemented between a coupled En-Var and an atmospheric 4D-Var: in that case, the global increment, used as input for the Jo computation for example, would be the ocean increment coming from the En-Var, while the atmospheric increment is the sum of the En-Var and 4D-Var increments. This can be interpreted as a generalisation of the concept of balanced and unbalanced components where here one partial increment from En-Var is "balanced" with the ocean, the atmospheric 4D-Var adding an "unbalanced" part. In the atmospheric part of the domain, 4D-Var would expand on the sub-space of ensemble perturbations brought by the En-Var and retain the quality of the 4D-Var analysis while the En-Var brings the benefits of coupling to the atmospheric analysis from an algorithm that is potentially easier to develop (no need for tangent linear and adjoint of the coupler and no explicit coupling term in the cost function). To continue the analogy with balanced and unbalanced control variables, the background error term for each part of the atmospheric increment would have to be adjusted to account for the fact each is a partial increment. The increment from the En-Var would be subtracted from the full 4D-Var increment to leave an unbalanced control variable. Such a coupled system would also benefit from all the flexibility of common variational methods such as bias correction, correlated observation error handling or additional constraints like Jc-DFI.

5.4   Perspectives

Coupled data assimilation is a relatively new field of research. Many questions are still open and future directions might evolve as experience is gained at ECMWF and in the scientific community. For example, errors in the systems to be coupled in the Earth system modelling context typically evolve with different characteristic time scales. This would indicate a potential advantage of using assimilation windows of different length for the different sub-systems, though it is not clear at this point how crucial this aspect is in realistic coupled DA systems (e.g., Lea, Mirouze, Martin, King, Hines, Walters, and Thurlow 2015). At present it is only possible to run the two sub-systems with the same window length. Therefore if the window length is optimised for one sub-system it may be sub-optimal for the other. The performance of using window lengths of 12 and 24 hours for both sub-systems will be studied using the CERA coupled reanalysis system using a full observing system. If it is found the trade-off in window length leads to significant sub-optimality in either sub-system further research into using different window lengths in different sub-systems will be needed.

Biases between the different components of the Earth system model are another important aspect that might prove very important in practice. Biases are the main reason that prevented fully coupled assimilation of ozone in the current atmospheric DA system and it is very likely that bias correction would greatly influence SST analysis in a coupled data assimilation system. Research efforts towards observation and possibly models bias correction in the coupled data assimilation context will be necessary.

Although most of the discussion here focused on coupled data assimilation for the ocean and in a variational context, research will also continue towards a stronger coupling of the land surface analysis with the atmospheric 4D-Var. Finally, the real time availability of data for non-atmospheric components of the Earth system might also affect operational use of coupled data assimilation for NWP.

Developing assimilation methods for each component of the Earth system in a common framework such as OOPS will open the door to many possibilities for coupling, even if it is not the same variational method. It brings much more benefits than just the re-use of a large part of the standard 4D-Var code base and low maintenance costs.

6   Conclusions

In conclusion the DA strategy at ECMWF can be summarised in five broad areas:

The OOPS project

The Object Oriented Prediction System (OOPS) project is now making very good progress in collaboration with partners from Member States, notably Météo-France and the ALADIN/HIRLAM communities. This project creates a framework that will allow continued evolution of the 4D-Var system (e.g. weak constraint, saddle-point algorithm) and allow research across many Earth System components using a wide range of algorithms (e.g. 4D-En-Var, EnKF, and the extended α control variable). Therefore the development of a fully functional atmosphere 4D-Var system within OOPS remains the highest priority to allow OOPS to be evaluated as the future operational system, as well as porting of the EnKF code and the integration of the En-Var code by Météo-France. In the meantime new science

will continue to be evaluated using toy models in the OOPS framework, as has been done for example for the saddle-point method. A major benefit of OOPS will be to allow more research collaboration with partners in the Member States, as well as providing new operational capability at ECMWF.

Development of Hybrid Ensemble DA

The development of ECMWF's EDA system has enabled a hybrid ensemble and variational approach without the need to support a separate data assimilation system for the ensemble component. This highly successful approach will continue to be the ECMWF ensemble data assimilation system for the foreseeable future. However recognising that the cost of EDA is high, various options have been evaluated to understand more deeply the advantages and disadvantages of different approaches to the handling of ensemble information. This has involved the testing of alternative configurations of EDA, as well as the development of a separate ensemble system (in our case EnKF) and multi-analysis techniques (Hybrid Gain). Whilst the EnKF is not considered to be a candidate for a future operational system this research has highlighted directions for improved use of ensemble information (e.g. in the all-sky assimilation framework). Developing a lower cost but equally effective EDA remains the second highest priority after the OOPS project. A more cost-effective EDA will allow the strategy towards a seamless approach to ensemble data assimilation and forecasting to advance, in particular by enabling more EDA members.

4D-Var development

The re-introduction of a mean model error term in the stratosphere using a weak constraint approach with an improved estimate of Q has been an important step towards a more flexible Data Assimilation system. Given the improvements in B modelling and the challenges of implementing a global weak constraint for mean and random model error it remains uncertain how large the impact will be for different window lengths, but we still expect benefits for our 12 hour window. We will therefore continue to work on the weak constraint as a high priority. Characterising model uncertainty is a very active area of research and development, described in more detail in the 2016 SAC Special Topic paper on model uncertainty (Leutbecher et al., 2016). So the development of weak constraint needs to continue to take account of the latest developments in ensemble forecasting, as well as feeding into them.

The development of a framework for overlapping 12 hour 4D-Var windows could be extended to a more radical redesign, with overlapping windows being run very frequently, for example hourly. Such a quasi-continuous 4D-DA system could lift the DA system out of the critical path for the IFS, with the forecast being run from the latest analysis available at the time. This concept and others related to the configuration of 4D-Var, including how to run EDA in time for the ensemble forecasting within this or other candidate frameworks, will also be a major activity.

A third critical area of activity for 4D-Var will be the implementation of the saddle-point algorithm. Once the OOPS system is running 4D-Var, evaluation of saddle-point will begin, with a view to operational implementation subject to satisfactory results when OOPS is ready to go operational. This will make 4D-Var scalable for the foreseeable future and is therefore critical to its long term viability.

To make this new science possible it will also be crucial to continue maintaining and further developing the linearised model and its adjoint. This will also allow exploring increasing the analysis resolution

and, as importantly, will provide the critical modelling infrastructure for the assimilation of new classes of observations (e.g., space-born radar and lidar, lighting observations).

Coupled Data Assimilation

A key part of our DA strategy is to introduce more coupling. The CERA system has demonstrated successful ocean-atmosphere coupled data assimilation. The DA methodology for the Coupled components is of course an area of active research and will be developed in incremental stages beginning with the CERA system and exploring how best to develop this. In order to deliver this the development of ECMWF's own SST and sea ice data assimilation capability is necessary, as a key step towards the coupled DA. For the reasons presented in this paper in the longer term the 4D-En-Var approach for the coupled components appears to have many attractive features, though as discussed this does not exclude maintaining 4D-Var for the atmosphere. In this context we will benefit from the integration of 4D-En-Var into OOPS by Météo-France, and will therefore be able to easily undertake research in this area within the OOPS framework. In particular this can be investigated in the ocean data assimilation context, but in due course will extend to other coupled components (waves, sea ice, land surface). We anticipate active collaboration with partners in the Member States in this area.

Use of more observations

Another goal of our DA strategy is to use more fully the observations available to ECMWF. At present some observations are not used because in a single minimisation they add little, either because correlations are not well understood, or because there is redundancy in the Global Observing System. It has been proposed to study using different observation subsets, both in the context of perturbed members of the ensemble DA system, and perhaps also for overlapping DA windows. This has to be done carefully, especially when observation thinning relates more to unknown error correlations than to redundant information. However the idea to use the very large number of currently unused observations is attractive.

These five areas together form a single strategy. The OOPS development is a complex project, involving major changes and modernisation to almost every part of the IFS system. Such a large project takes many years to complete but very significant progress has been made. OOPS is an enabling project, from which many areas of new science can flow. These include critical areas for 4D-Var, such as weak constraint and time parallelisation, as well as the provision of other algorithms, for example 4DEnVar, that could be useful in contexts such as ensemble or coupled data assimilation. Coupled DA forms a key part of ECMWF's DA strategy now, in line with similar developments in coupled models.

The configuration of the hybrid component is critically dependent on the choice of DA algorithm. The EDA of 4D-Vars has proven to give outstanding performance. Ideas have been described aiming to deliver this same performance but at lower cost, to enable a larger ensemble which will benefit both the high resolution 4D-Var and ensemble forecasting. At present best results retain the 4D-Var component and therefore improvement of 4D-Var remains a key component of the hybrid DA approach.

In all the DA developments the question "can we use more observations?" is continually asked. It was shown how the DA developments, in 4D-Var, coupled DA and the hybrid methods, can enable more complete extraction of the information in the observations we have, and will have in future. In this

respect the opportunities arising from coupled DA, especially for surface and near-surface observations, is of particular interest.

Exploiting the synergies between the developments in 4D-Var, hybrid DA, coupled DA and improved use of observations is critical towards achieving the goals of the strategy.

Acknowledgements

REFERENCES

Ades, M., and P. J. van Leeuwen, 2015: The effect of the equivalent-weights particle filter on dynamical balance in a primitive equation model. *Mon. Weather. Rev.* **143**: 581–596.

Ajgl, J., M. Šimandl and J. Dunik, 2009: Millman's formula in data fusion. In *Proceedings of the 10th International PhD Workshop on Systems and Control*: 1–6. UTIA: Prague, Czech Republic.

Auligné, T., B. Ménétrier, A. C. Lorenc, M. Buehner, 2016: Ensemble-Variational Integrated Localized data assimilation. *Mon. Weather Rev.,* **144**, 3677-3696, doi: http://dx.doi.org/10.1175/MWR-D-15-0252.1

Anderson, J. L., 2001: An Ensemble Adjustment Kalman Filter for data assimilation. *Mon. Weather Rev*. **129**: 2884–2903.

Anderson, J., T. Hoar, K. Raeder, H. Liu, N. Collins, R. Torn, and A. Avellano, 2009: The Data Assimilation Research Testbed: a community facility. Bull. Am. Meteorol. Soc., **90**: 1283–1296, doi: 10.1175/2009BAMS2618.1

Andersson, E., Pailleux, J., Thépaut, J.-N., Eyre, J. R., McNally, A. P., Kelly, G. A. and Courtier, P., 1994: Use of cloud-cleared radiances in three/four-dimensional variational data assimilation. *Q. J. R. Meteorol. Soc.* **120**: 627–653. doi:10.1002/qj.49712051707

Andersson, E., Haseler, J., Undén, P., Courtier, P., Kelly, G., Vasiljevic, D., Brankovic, C., Gaffard, C., Hollingsworth, A., Jakob, C., Janssen, P., Klinker, E., Lanzinger, A., Miller, M., Rabier, F., Simmons, A., Strauss, B., Viterbo, P., Cardinali, C. and Thépaut, J.-N., 1998: The ECMWF implementation of three-dimensional variational assimilation (3D-Var). III: Experimental results. *Q. J. R. Meteorol. Soc.* **124**: 1831–1860. doi:10.1002/qj.49712455004

Andersson, E., M. Fisher, R. Munro and A. McNally, 1999: Diagnosis of Background Errors for Radiances and Other Observable Quantities in a Variational Data Assimilation Scheme, and the Explanation of a Case of Poor Convergence*, Technical Memorandum* 296. ECMWF: Reading, UK.

Bauer, P., A. J. Geer, P. Lopez and D. Salmond, 2010: Direct 4D-Var assimilation of all-sky radiances. Part I: Implementation. *Q. J. R. Meteorol. Soc.* **136**: 1868–1885. doi:10.1002/qj.659

Belo Pereira, M. and L. Berre, 2006: The Use of an Ensemble Approach to Study the Background Error Covariances in a Global NWP Model. *Mon. Wea. Rev.*, **134**, 2466–2489, doi: 10.1175/MWR3189.1.

Berre, L., O. Pannekoucke, G. Desroziers, S. E. Stefanescu, B. Chapnik and L. Raynaud, 2007: A variational assimilation ensemble and the spatial filtering of its error covariances: increase of sample size by local spatial averaging. In *Proceedings of Workshop on Flow-dependent Aspects of Data Assimilation*: 151–168. ECMWF: Reading, UK.

Berre, L., S. E. Stefanescu and M. Belo Pereira, 2006: The representation of the analysis effect in three error simulation technique. *Tellus* **58A**: 196–209.

Bishop, C., 2016: Using trajectories of ensemble analyses and tools from weak constraint 4D-Var to represent model uncertainty. In *ECMWF/WWRP Workshop on Model Uncertainty*, available at http://www.ecmwf.int/en/learning/workshops-and-seminars/wwrp-workshop-model-uncertainty

Bonavita, M., L. Raynaud and L. Isaksen, 2011: Estimating background-error variances with the ECMWF ensemble of data assimilation system: The effect of ensemble size and day-to-day variability. *Q. J. R. Meteorol. Soc.,* **137**: 423–434.

Bonavita, M., L. Isaksen and E. V. Hólm, 2012: On the use of EDA background-error variances in the ECMWF 4D-Var. *Q. J. R. Meteorol. Soc.,* **138**: 1540–1559.

Bonavita, M., M. Hamrud and L. Isaksen, 2015: EnKF and Hybrid Gain Ensemble Data Assimilation. Part II: EnKF and Hybrid Gain results. Mon. Weather Rev., doi: http://dx.doi.org/10.1175/MWR-D-15-0071.1

Bonavita, M., E. Hólm, L. Isaksen and M. Fisher, 2016: The evolution of the ECMWF hybrid data assimilation system. *Q. J. R. Meteorol. Soc.* **142**: 287–303. doi:10.1002/qj.2652

Bowler, N. E., J. Flowerdew and S. R. Pring, 2013: Tests of different flavours of EnKF on a simple model. *Q. J. R. Meteorol. Soc.* **139**: 1505–1519. doi:10.1002/qj.2055

Bowler, N. E., A. M. Clayton, M. Jardak, E. Lee, A. C. Lorenc, C. Piccolo, S. R. Pring, M. A. Wlasak, D. M. Barker, G. W. Inverarity and R. Swinbank, 2016a: The development of an ensemble of 4D-ensemble variational assimilations. *Q. J. R. Meteorol. Soc.*, under review.

Bowler, N. E., A. M. Clayton, M. Jardak, E. Lee, A. C. Lorenc, C. Piccolo, S. R. Pring, M. A. Wlasak, D. M. Barker, G. W. Inverarity and R. Swinbank, 2016b: The development of a hybrid 4D-ensemble variational assimilation system. *Q. J. R. Meteorol. Soc.,* under review.

Bowler, N. E., 2016: On the diagnosis of model error statistics using weak-constraint data assimilation, *Q. J. R. Meteorol. Soc.,* submitted.

Browne, P. A., and P. J. van Leeuwen, 2015: Twin experiments with the equivalent weights particle filter and HadCM3. *Q. J. R. Meteorol. Soc*. **141**: 3399–3414. doi:10.1002/qj.262

Buehner, M., R. McTaggart-Cowan, A. Beaulne, C. Charette, L. Garand, S. Heilliette, E. Lapalme, S. Laroche, S. R. Macpherson, J. Morneau and A. Zadra, 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., R. McTaggart-Cowan, S. Heilliette, 2016: An Ensemble Kalman Filter for NWP based on Variational Data Assimilation: VarEnKF. Presentation at "*Blueprints for Next-Generation Data Assimilation Systems*", 8-10 March 2016; Boulder, Colorado, USA. https://www.da.ucar.edu/sites/default/files/Buehner_NCAR_DA_Blueprint_VarEnKF.pdf

Buehner, M., R. McTaggart-Cowan, S. Heilliette, 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, in press.

Buizza, R., 1993: Impact of a simple vertical diffusion scheme and of the optimisation time interval on optimal unstable structures, *Technical Memorandum* 192. ECMWF: Reading, UK.

Buizza, R., M. Leutbecher, and L. Isaksen, 2008: Potential use of an ensemble of analyses in the ECMWF Ensemble Prediction System. *Q. J. R. Meteorol. Soc.* **134**: 2051–2066.doi:10.1002/qj.346

Burgers, G., P. J. van Leeuwen and G. Evensen, 1998: On the analysis scheme in the Ensemble Kalman Filter. *Mon. Weather Rev*. **126**: 1719–1724.

Chapelle, D., M. Fragu, V. Mallet and P. Moireau, 2013: Fundamental principles of data assimilation underlying the Verdandi library: applications to biophysical model personalization within euHeart. *Medical and Biological Engineering and Computing* **51**:1221–1233.

Clayton, A. M., A. G. Lorenc and D. M. Barker, 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

Courtier, P., J.-N. Thépaut and A. Hollingsworth, 1994: A strategy for operational implementation of 4D-Var using an incremental approach. *Q. J. R. Meteorol. Soc.,* **120**: 1367–1387.

Cullen, M. J. P., 2010: A demonstration of cycled 4D-Var in the presence of model error. *Q. J. R. Meteorol. Soc.* **136**: 1379–1395, doi:10.1002/qj.653.

Cullen, M. J. P., 2013: Analysis of cycled 4D-Var with model error. *Q. J. R. Meteorol. Soc.***139**: 1473–1480, doi:10.1002/qj.2045.

Dee, D. P., 2004: Variational bias correction of radiance data in the ECMWF system. In *Proceedings of Workshop on Assimilation of High Spectral Resolution Sounders in NWP*: 97–112. ECMWF: Reading, UK.

de Rosnay, P., M. Drusch, D. Vasiljevic, G. Balsamo, C. Albergel, and L. Isaksen, 2013: A simplified Extended Kalman Filter for the global operational soil moisture analysis at ECMWF, *Q. J. R. Meteorol. Soc.* **139**: 1199–1213.

Desroziers, G., 1997: A coordinate change for data assimilation in spherical geometry of frontal structures. *Mon. Weather Rev*. **125**: 3030–3038.

ECMWF, SAC paper: ECMWF/SAC/40(11)5, 2011: Weak-Constraint and Long-Window 4D-Var.

ECMWF, SAC paper: ECMWF/SAC/40(11)7, 2011: Data assimilation vision for 2014-2018.

ECMWF, SAC paper: ECMWF/SAC/45(16)8, 2016: On stochastic representations of model uncertainties.

ECMWF, SAC paper: ECMWF/SAC/45(16)9, 2016: A strategy for hybrid data assimilation.

Evensen, G., 1994: Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *J. Geophys. Res.* **99**(C5): 10143–10162.

Fisher, M., 1998: *Development of a Simplified Kalman Filter, Technical Memorandum* 260. ECMWF: Reading, UK.

Fisher, M., and E. Andersson, 2001*: Developments in 4D-Var and Kalman Filtering, Technical Memorandum* 347. ECMWF: Reading, UK.

Fisher, M., M. Leutbecher, and G. A. Kelly, 2005: On the equivalence between Kalman smoothing and weak-constraint four-dimensional variational data assimilation. *Q. J. R. Meteorol. Soc.* **131**: 3235–3246, doi:10.1256/qj.04.142.

Fisher, M., 2007: The sensitivity of analysis errors to the specification of background error covariances. *Proceedings of Workshop on Flow-dependent Aspects of Data Assimilation*: 27–36. ECMWF: Reading, UK.

Fisher, M., S. Gratton, S. Gürol, Y. Trémolet and X. Vasseur, 2016: Low rank updates in preconditioning the saddle-point systems arising from data assimilation problems. *Optimization Methods and Softwar*e. http://dx.doi.org/10.1080/10556788.2016.1264398

Fisher, M., and S., Gürol, 2017: Parallelisation in the Time Dimension of Four-Dimensional Variational Data Assimilation. *Q. J. R. Meteorol. Soc.* accepted author manuscript. doi:10.1002/qj.2997

Flowerdew, J., 2015: Towards a theory of optimal localisation. *Tellus* **67A**, 25257, http://dx.doi.org/10.3402/tellusa.v67.25257

Gauthier, P., and J.-N. Thépaut, 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.

Geer, A. J., F. Baordo, N. Bormann and S. English, 2014: All-sky assimilation of microwave humidity sounders, *Technical Memorandum* 741. ECMWF: Reading, UK.

Giles, M. B., (2008): Multilevel Monte Carlo path simulation. *Oper. Res*. **56**(3): 607–617.

Gregory, A., C. J. Cotter and S. Reich, 2016: Multilevel ensemble transform particle filtering. *SIAM J. Sci. Comput.* **30**(3): A1317–A1338.

Hamrud, M., M. Bonavita, and L. Isaksen, 2015: EnKF and hybrid gain ensemble data assimilation. Part I: EnKF implementation. *Mon. Weather Rev*. **143**: 4847−4864.

Haslehner, M., T. Janjic and G. C. Craig, 2016: Testing particle filters on simple convective-scale models. Part 2: A modified shallow-water model. *Q. J. R. Meteorol. Soc.,* **142**: 1628–1646.

He, H., and E. A. Garcia, 2009: Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering* **21**(9): 1263–1284.

Houtekamer, P. L., and H. L. Mitchell, 1998: Data assimilation using an ensemble Kalman filter technique. *Mon. Weather Rev*. **126**: 796–811.

Houtekamer, P. L., and F. Zhang, 2016: Review of ensemble Kalman filter for atmospheric data assimilation. *Mon. Weather Rev*. htpp://dx.doi.org/10.1175/MWR-D-15-0440.1

Hunt, B. R., E. J. Kostelich and I. Szunyogh, 2007: Efficient data assimilation for spatiotemporal chaos: a local ensemble transform Kalman filter. *Physica D* **230**: 112–126.

Isaksen, L., M. Fisher and J. Berner, 2007: Use of analysis ensembles in estimating flow-dependent background error variance. In *Proceedings of the Workshop on Flow-dependent Aspects of Data Assimilation*: 65–86. ECMWF: Reading, UK.

Isaksen, L., M. Bonavita, R. Buizza, M. Fisher, J. Haseler, M. Leutbecher and L. Raynaud, 2010: *Ensemble of data assimilations at ECMWF, Technical Memorandum* 636. ECMWF: Reading, UK.

Janisková, M., and P. Lopez, 2013: Linearized physics for data assimilation at ECMWF. In: S.K. Park and L. Xu (eds.), *Data assimilation for Atmospheric, Oceanic and Hydrological Applications* (Vol. II), doi: 10.1007/978-3-642-35088-7, ©Springer-Verlag Berlin Heidelberg, 251-286.

Janisková, M., 2015: Assimilation of cloud information from space-borne radar and lidar: experimental study using a 1D+4D-Var technique, *Q. J. R. Meteorol. Soc.* **141**: 2708–2725.

Janisková, M. and C. Cardinali, 2017: On the impact of the diabatic component in the Forecast Sensitivity Observation Impact diagnostics. *Data assimilation for Atmospheric, Oceanic and Hydrological Applications* (Vol. III), doi: 10.1007/978-3-319-43415-5_22, ©Springer-Verlag Berlin Heidelberg, 483–511.

Järvinen, H., J.-N. Thépaut and P. Courtier, 1996: Quasi-continuous variational data assimilation. *Q. J. R. Meteorol. Soc.* **122**: 515–534. doi:10.1002/qj.49712253011

Kucukkaraca, E., and M. Fisher, 2006: Use of Analysis Ensembles in Estimating Flow-dependent Background Error Variances. *Technical Memorandum* 492. ECMWF: Reading, UK.

Laloyaux, P., M. Balmaseda, D. Dee, K. Mogensen and P. Janssen, 2016: A coupled data assimilation system for climate reanalysis. *Q. J. R. Meteorol. Soc.* **142**: 65–78. doi:10.1002/qj.2629

Lang, S. T. K., M. Bonavita and M. Leutbecher, 2015: On the impact of re-centring initial conditions for ensemble forecasts. *Q. J. R. Meteorol. Soc.* **141**: 2571–2581. doi:10.1002/qj.2543

Lea, D. J., I. Mirouze, M. J. Martin, R. R. King, A. Hines, D. Walters and M. Thurlow, 2015: Assessing a new coupled data assimilation system based on the Met Office coupled atmosphere–land–ocean–sea ice model. *Mon. Weather Rev.* **143**: 4678–4694, doi: 10.1175/MWR-D-15-0174.1

Le Dimet, F.-X., and O. Talagrand, 1986: Variational algorithms for analysis and assimilation of meteorological observations: theoretical aspects. *Tellus* **38A**: 97–110. doi:10.1111/j.1600-0870.1986.tb00459.x

Legrand, R., and Y. Michel, 2014: Modelling background error correlations with spatial deformations: a case study. *Tellus* **66A**: 23984, http://dx.doi.org/10.3402/tellusa.v66.23984

Lei, L., and J. S. Whitaker, 2015: Model space localization is not always better than observation space localization for assimilation of satellite radiances. *Mon. Weather Rev*. **143**: 3948–3955, doi: 10.1175/MWR-D-14-00413.1.

Leutbecher, M., and T. N. Palmer, 2008: Ensemble forecasting. *J. Comput. Phys*. **227**: 3515–3539.

Leutbecher M., S-J. Lock, P. Ollinaho, S. T. K. Lang, G. Balsamo, P. Bechtold, M. Bonavita, H.M. Christensen, M. Diamantakis, E. Dutra, S. English, M. Fisher, R. M. Forbes, J. Goddard, T. Haiden, R. J. Hogan, S. Juricke, H. Lawrence, D. MacLeod, L. Magnusson, S. Malardel, S. Massart, I. Sandu, P. K. Smolarkiewicz, A. Subramanian, F. Vitart, N. Wedi, A. Weisheimer, 2016: Stochastic representations of model uncertainties at ECMWF: State of the art and future vision. ECMWF Tech Memo 785.

Liu, C., Q. Xiao and B. Wang, 2008: An ensemble-based four-dimensional variational data assimilation scheme. Part I: technical formulation and preliminary test. *Mon. Weather Rev.* **136**(9):3363–3373.

Lopez, P., and E. Moreau, 2005: A convection scheme for data assimilation: Description and initial tests, *Q. J. R. Meteorol. Soc.* **131**: 409–436.

Lopez, P., 2011: Direct 4D-Var Assimilation of NCEP Stage IV Radar and Gauge Precipitation Data at ECMWF, *Mon. Weather Rev.* **139**: 2098–2116.

Lopez, P., 2016: A lightning parameterization for the ECMWF Integrated Forecasting System, *Mon. Weather Rev.* **144**: 3057–3075.

Lorenc, A. C., 2003: The potential of the ensemble Kalman filter for NWP: A Comparison with 4D-Var. *Q. J. R. Meteorol. Soc.* **129**: 3183–3203.

Lorenc, A.C., M. Jardak, T.J. Payne, N.E. Bowler and M. Wlasak, 2016: Computing an ensemble of variational data assimilations using its mean and perturbations. *Q. J. R. Meteorol. Soc.,* accepted author manuscript doi:10.1002/qj.2965.

Mahfouf, J.-F., 1999: Influence of physical processes on the tangent-linear approximation, *Tellus* **51A**: 147–166.

Mahmoud, M. S., and H. M. Khalid, 2013: Distributed Kalman filtering: a bibliographic review. *IET Control Theory and Applications* **7**(4): 483–501.

Ménard, R., and R. Daley, 1996: The application of Kalman smoother theory to the estimation of 4DVAR error statistics. *Tellus* **48A**: 221–237. doi:10.1034/j.1600-0870.1996.t01-1-00003.x

Ménétrier, B., and T. Auligné, 2015: Optimized localization and hybridization to filter ensemble-based covariances. *Mon. Weather Rev.* **143**: 3931–3947, doi: 10.1175/MWR-D-15-0057.1.

Meunier, L-F, S. English and P. Janssen, 2014: Improved Ocean Emissivity Modelling for Assimilation of Microwave Imagers using Foam Coverage Derived from a Wave Model. *Report of Visiting Scientist mission NWP_VS12_02*. Document NWPSAF-MO-VS-049, available from nwpsaf.eu.

Michel, Y., 2013a: Estimating deformations of random processes for correlation modelling in a limited area model. *Q. J. R. Meteorol. Soc*. **139**: 534–547.

Michel, Y., 2013b: Estimating deformations of random processes for correlation modelling: methodology and the one-dimensional case. *Q. J. R. Meteorol. Soc.*, **139**: 771–783.

NCEP, 2016: *History of Recent Modifications to the Global Forecast/Analysis System*. Available at http://www.emc.ncep.noaa.gov/gmb/STATS/html/model_changes.html

Orr, A., P. Bechtold, J. Scinocca, M. Ern and M. Janisková, 2010: Improved middle atmosphere climate and forecasts in the ECMWF model through a nonorographic gravity wave drag parameterization. *J. Atmos. Sci*. **23**: 5905–5926.

Pannekoucke, O., E. Emili and O. Thual, 2014: Modelling of local length-scale dynamics and isotropizing deformations. *Q. J. R. Meteorol. Soc.* **140**: 1387–1398. doi:10.1002/qj.2204

Pellerej, R., A. Vidard and F. Lemarié, 2016: Toward Variational Data Assimilation for Coupled Models: First Experiments on a Diffusion Problem. CARI, Tunis, October 2016, http://www.cari-info.org/Actes-2016/

Penny, G. S., 2014: The Hybrid Local Ensemble Transform Kalman Filter. *Mon. Weather Rev.* **142**: 2139–2149. doi:http://dx.doi.org/10.1175/MWR-D-13-00131.1

Piccolo, C. and M. Cullen, 2012: A new implementation of the adaptive mesh transform in the Met Office 3D-Var System. *Q. J. R. Meteorol. Soc.* **138**: 1560–1570. doi:10.1002/qj.1880

Pires, C., R. Vautard and O. Talagrand, 1996: On extending the limits of variational assimilation in nonlinear chaotic systems. *Tellus* **48A**: 96–121. doi:10.1034/j.1600-0870.1996.00006.x

Poterjoy, J., and J. L. Anderson, 2016: Efficient assimilation of simulated observations in a high-dimensional geophysical system using a localized particle filter. *Mon. Weather Rev.* **144**: 2007–2020. doi:10.1175/MWR-D-15-0322.1.

Rabier, F., H. Järvinen, E. Klinker, J.-F. Mahfouf and A. Simmons, 2000: The ECMWF operational implementation of four-dimensional variational assimilation. Part I: Experimental results with simplified physics. *Q. J. R. Meteorol. Soc.* **126**: 1143–1170.

Rainwater, S., and B. Hunt, 2013: Mixed-resolution ensemble data assimilation. *Mon. Weather Rev.* **141**: 3007–3021.

Reich, S., 2013: A nonparametric ensemble transform method for Bayesian inference. *SIAM J. Sci. Comput.* **35**: A2013–A2024.

Robert, S., and H. R. Kuensch, 2016: Local ensemble Kalman particle filters for efficient data assimilation. *arXiv*:1605.05476v1 [stat.AP] 18 May 2016.

Rodwell, M. J., S. Lang, N. B. Ingleby, N. Bormann, E. Hólm, F. Rabier, D. S. Richardson and M. Yamaguchi, 2016: Reliability in ensemble data assimilation. *Q. J. R. Meteorol. Soc.* **142**: 443–454. doi:10.1002/qj.2663

Shin, V., Y. Lee and T.-S. Choi, 2006: Generalized Millman's formula and its application for estimation problems. *Signal Processing* **86**: 257–266.

Slivinski, L., and C. Snyder, 2016: Exploring practical estimates of the ensemble size necessary for particle filters. *Mon. Weather Rev.* **144**: 861–875.

Smith, P., A. Fowler, and A. Lawless, 2015: Exploring strategies for coupled 4D-Var data assimilation using an idealised atmosphere-ocean model. *Tellus* **67A**: 27025, http://dx.doi.org/10.3402/tellusa.v67.27025

Tippett, M. K., J. L. Anderson, C. H. Bishop, T. M. Hamill and J. S. Whitaker, 2003: Ensemble square root filters. *Mon. Weather Rev*. **131**:1485–1490.

Todling, R., 2015: A lag-1 smoother approach to system error estimation: Sequential method. *Q. J. R. Meteorol. Soc.* **141**: 1502–1513, doi:10.1002/qj.2460.

Tompkins, A. M., and M. Janisková, 2004: A cloud scheme for data assimilation: Description and initial tests. *Q. J. R. Meteorol. Soc.* **130**: 2495–2517. doi:10.1256/qj.03.162

Trémolet, Y., 2006: Accounting for an imperfect model in 4D-Var. *Q. J. R. Meteorol. Soc.* **132**: 2483–2504. doi:10.1256/qj.05.224

Trémolet, Y., 2008: Computation of observation sensitivity and observation impact in incremental variational data assimilation. *Tellus* **60A**:964–978.

van Leeuwen, P. J., 2015: Nonlinear data assimilation for high-dimensional systems. In *Frontiers in Applied Dynamical Systems: Reviews and Tutorials*, Volume 2, pp. 1-73. Springer-Verlag, New York.

Wang, X., C. Snyder, and T. M. Hamill, 2007: On the Theoretical Equivalence of Differently Proposed Ensemble 3D-Var Hybrid Analysis Schemes. *Mon. Weather Rev.* **135**(1):222–227

Wang, X., D. M. Barker, C. Snyder and T. M. Hamill, 2008: A Hybrid ETKF–3D-Var Data Assimilation Scheme for the WRF Model. Part I: Observing System Simulation Experiment. *Mon. Weather Rev.* **136**: 5116–5131.

Whitaker, J.S. and T. Hamill, 2002: Ensemble data assimilation without perturbed observations. *Mon. Weather Rev.* **130**: 1913–1924. http://dx.doi.org/10.1175/1520-0493(2002)130<1913:EDAWPO>2.0.CO;2

Whitaker, J.S., and L. Lei, 2016: EnKF radiance assimilation. Presentation at "*The 7th EnKF Data Assimilation Workshop*", 23-27 May 2016, State College, PA, USA. Available at http://adapt.psu.edu/2016EnKFWorkshop

Yashiro, H., K. Terasaki, T. Miyoshi and H. Tomita, 2016: Performance evaluation of throughput-aware framework for ensemble data assimilation: The case of NICAM-LETKF. *Geosci. Model Dev.* **9**:2293–2300, doi:10.5194/gmd-9-2293-2016.

Zhu, M., P. J. van Leeuwen and J. Amezcua, 2016: Implicit equal-weights particle filter. *Q. J. R. Meteorol. Soc.* **142**: 1904–1919.