Technical Memo

ECMWF
European Centre for Medium-Range Weather Forecasts

856

A new hybrid formulation for the background error covariance in the IFS: evaluation

S. Massart
(Research Department)

December 2019

Series: ECMWF Technical Memoranda

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

Contact: library@ecmwf.int

Summary

A new hybrid formulation for the background errors in the European Centre for Medium-Range Weather Forecasts (ECMWF) 4D-Var was proposed in Massart (2018) and tested in simple configurations. In the following document, we evaluate this new formulation in a more realistic configuration. In order to perform the evaluation, the new formulation had to be adapted to the current implementation of ECMWF 4D-Var. The 4D-Var analysis is the result of an incremental 4D-Var that successively minimises linear versions of the cost function (outer loop iteration). In ECMWF incremental 4D-Var implementation, the resolution of the tangent-linear and adjoint versions of the model and observation operators used in the analysis increases with each outer loop. We demonstrate that the new hybrid formulation can be used together with this implementation of the incremental 4D-Var, but one has to consider carefully the change of resolution between outer loops.

We designed three experiments to carry out the evaluation. The first experiment is based on static background errors and referred to as $\mathbf{B}^s$. The second experiment is based on the operational formulation of the background errors as implemented in CY45R1 and referred to as $\mathbf{B}^o$. The last experiment is based on the new hybrid formulation with an hybrid weight of 50 %, and is referred to as α-$\mathbf{B}^h$. They are all designed so that the experiments have the same background error variances on average.

The evaluation is based on a comparison between the α-$\mathbf{B}^h$ experiment and the two other at the horizontal resolution of TCo 399, with 3 outer loops, and over a winter season, from November 2017 to February 2018 (4 months). The comparison against the static formulation aims at assessing the impact of flow-dependent background errors. The comparison against CY45R1 operational formulation aims at assessing if the new formulation could complement the current one and help improving the forecast scores.

The main result from the evaluation is that the analysis of the α-$\mathbf{B}^h$ experiment is spatially smoother that the analysis from the two other experiments even if, for this experiment, the fields of the atmospheric model variables have more energy in the small scales (from the wavenumber around 100). We deduce that this feature accelerates the convergence of the minimisation of the α-$\mathbf{B}^h$ experiment. This experiment requires between 1 and 2 less iterations than for the $\mathbf{B}^s$ experiment, and between 5 and 7 less iterations than for the $\mathbf{B}^o$ experiment. We also deduce that the smoothness of the analysis from the α-$\mathbf{B}^h$ experiment is likely to be the reason why the first-guess derived from it is on average closer to the observations than the one derived from the two other experiments, by 0.5 % compared to $\mathbf{B}^s$ and by 1.3 % compared to $\mathbf{B}^o$. For the satellite sounding data, the $\mathbf{B}^o$ experiment performs better and the main difference is for the ATMS instrument for which the $\mathbf{B}^o$ first-guess is closer to the observations by an average of 3.3 % compared to α-$\mathbf{B}^h$ first-guess. This difference may explain why the forecast (compared against its own analysis or observations) from the α-$\mathbf{B}^h$ experiment is the best, but only up to day 3 to 5. Then, the medium-range forecast from the $\mathbf{B}^o$ experiment tends to be the best.

Overall, the results show that the new hybrid formulation of the background errors provides better forecast scores than when using static background errors. Yet, it currently fails to compete with the current formulation for medium-range forecasts. Nonetheless, the results also show that the current formulation could be improved. For example, the wind forecast is better for all forecast ranges in the southern hemisphere and tropics when it is derived from the α-$\mathbf{B}^h$ experiment. The relative humidity forecast from the α-$\mathbf{B}^h$ experiment is also in general the best. These results encourage us to further develop the new hybrid formulation.

Contents

1 Introduction **5**

2 Implication of the 4D-Var incremental formulation for the background errors modelling **6**

2.1 Current implementation of the incremental formulation 6

2.1.1 First guess states 6

2.1.2 Linearisation of the cost function 7

2.1.3 Change of variable 7

2.1.4 Incremental approach 8

2.2 Incremental formulation with the new hybrid background error model 9

2.2.1 Algorithm 9

2.2.2 Resolution of the perturbations 9

2.2.3 Localisation length-scale 10

3 Evaluation of the incremental formulation with a resolution-dependent localisation length-scale **11**

3.1 Single observation experiment 11

3.1.1 Experimental setup 12

3.1.2 Results 12

3.2 Full observing system experiment 12

3.2.1 Experimental setup 13

3.2.2 Results 14

4 Evaluation of the new formulation over a winter season **16**

4.1 Configuration 16

4.1.1 Experiments 16

4.1.2 Static background errors 17

4.1.3 Hybrid α-$\mathbf{B}^h$ background errors 17

4.1.4 Localisation of specific humidity perturbations 18

4.2 Results 19

4.2.1 Convergence diagnostics 19

4.2.2 Analysis 21

4.2.3 First-guess 24

4.2.4 Forecast 27
4.2.5 Interaction with VarBC 31

5 Conclusion **32**

6 Acknowledgements **33**

Appendix A Building a climatology of the background errors **35**

Appendix B Change in analysis mean fields **36**
B.1 Zonal mean 36
B.2 Mean spectra 37

Appendix C Statistics of the experiments against observations **38**
C.1 Analysis 38
C.1.1 Wind related observations 38
C.1.2 Temperature related observations 38
C.1.3 Humidity related observations 39
C.2 First-guess 39
C.2.1 Wind related observations 39
C.2.2 Temperature related observations 40
C.2.3 Humidity related observations 40
C.2.4 Satellite related observations 40

1 Introduction

The meteorological analysis of the European Centre for Medium-Range Weather Forecasts (ECMWF) aims at finding the best meteorological state at the beginning of a time period called the assimilation window, given all the available observations distributed over the assimilation window and given an *a priori* estimate (or background) of the meteorological state at the beginning of the assimilation window. The search of the optimal initial state also requires the background error and the observation error covariances (Talagrand, 1997). The background error covariances are of the order $\mathcal{O}\left(10^9\right)$. That makes them too large to be computed and stored in a matrix form. Instead, the background error covariances is decomposed in a sequence of operators (see Appendix C of Fisher et Andersson, 2001). Currently, each of the operators is parametrised using an ensemble of short range forecasts that are valid at the same time as the analysis and that are issued from an ensemble of data assimilations (EDA, Bonavita *et al.*, 2012). The background error correlation operator also uses a climatology of the background errors. This makes the present formulation of the background error flow dependent (using the EDA members of the day) and hybrid (using climatology as part of the correlation operator). We refer hereafter to this formulation as the *operational* formulation (or $\mathbf{B}^o$ formulation).

In Massart (2018), we described another approach to model the background error covariances. This new approach makes direct use of the ensemble of short range forecasts issued from the EDA members of the day to build an ensemble-based estimate of the background error covariances. However, this approach requires a localisation function to reduce the sampling noise of the ensemble-based estimate. The estimate is then combined with static background error covariances based on a climatology. By combining an ensemble-based estimate and static background error covariances, this approach is a hybrid formulation as the current formulation. There are several possible implementations of this new hybrid formulation (Desroziers *et al.*, 2014). In the IFS, we chose to augment the control variable with so-called alpha variables (Massart, 2018). We refer hereafter to this formulation of the background error covariances as the alpha control variable formulation (or α-$\mathbf{B}^h$ formulation).

In Massart (2018), the results from the α-$\mathbf{B}^h$ formulation were presented only for simple configurations designed to validate the correctness of its implementation. In this document we aim at a more complete evaluation of this new formulation. To this end, we had to enhance the implementation of the α-$\mathbf{B}^h$ formulation to account for the incremental approach currently used in the ECMWF 4D-Var analysis (Courtier *et al.*, 1994). The first part of this report describes the incremental formulation and the specifics for the α-$\mathbf{B}^h$ formulation implementation. One main feature of the incremental formulation is to perform a series of minimisations of a cost function, each sequence of the series being referred to as outer loop. For each outer loop, the cost function to be minimised is a linear version of the original cost function, but uses a tangent-linear and adjoint versions of the model and observation operators at a lower resolution than the original one. We show that the localisation length-scale used for the α-$\mathbf{B}^h$ formulation depends on the resolution used for the model in each outer loop. The second part of this document presents the impact of a resolution-dependent localisation length-scale on the analysis for a simple configuration with 3 outer loops. This helps setting up the configuration for a more complete evaluation of the α-$\mathbf{B}^h$ formulation over a 4 months period covering the 2017/2018 winter. This evaluation is presented in the last section before the general conclusion.

2 Implication of the 4D-Var incremental formulation for the background errors modelling

The implementation of the 4D-Var in the ECMWF Integrated Forecasting System (IFS) aims at estimating the best model state $\mathbf{x}(t_0)$ at the initial time t_0 of the assimilation window (spanning from time t_0 to time t_N) given all the information available. In its simplest form, the information comes from a set of observations $\{\mathbf{y}_i^o, \forall i \in [0,N]\}$ distributed over the $N+1$ time-slots t_0 to t_N, and from a prior estimate or background $\mathbf{x}^b$ of the initial state, together with their error covariances, respectively the $\{\mathbf{R}_i, \forall i \in [0,N]\}$ and $\mathbf{B}$ matrices. The best estimate referred to as $\mathbf{x}^a(t_0)$, is obtained by minimising the non-linear cost function

$$\begin{aligned} \mathscr{J}\left(\mathbf{x}(t_0)\right) &= \frac{1}{2}\left(\mathbf{x}(t_0)-\mathbf{x}^b\right)^T \mathbf{B}^{-1}\left(\mathbf{x}(t_0)-\mathbf{x}^b\right) \\ &+ \frac{1}{2}\sum_{i=0}^{N}\left[H_i \circ M_i\big(\mathbf{x}(t_0)\big)-\mathbf{y}_i^o\right]^T \mathbf{R}_i^{-1}\left[H_i \circ M_i\big(\mathbf{x}(t_0)\big)-\mathbf{y}_i^o\right], \end{aligned} \tag{1}$$

where H_i and M_i are respectively the observation operator at time t_i and the non-linear model that propagates the model state $\mathbf{x}(t_0)$ from initial time t_0 to time t_i.

The minimisation of the cost function of Eq. (1) is currently achieved by an iterative algorithm that computes successive values of the cost function and its gradient (inner loop iteration). The computational cost of an individual evaluation of the cost function and its gradient over a time window is about four times as expensive as a forecast over the same time window. As an example, for an assimilation window of $t_N - t_0 = 12$ hours, reaching the minimum of the cost function in 30 iterations would have the same computational cost as a 60 day forecast. This computational cost of the minimisation process being not affordable, Courtier *et al.* (1994) proposed the incremental formulation in order to reduce it. The incremental formulation consists in (i) a linearisation of the cost function around a first guess state, and (ii) a reduction of the resolution of the model used in the linear cost function obtained after step (i). The next sections present in more details the incremental formulation and the implementation of the α-$\mathbf{B}^h$ formulation when using the incremental formulation.

2.1 Current implementation of the incremental formulation

2.1.1 First guess states

Let us introduce a sequence of $K+1$ first guess states $\mathbf{x}_k^g$, where $k \in [0,K]$ indexes the linearisation loop of the non-linear cost function of Eq. (1). The initial first guess ($k=0$) is chosen to be the background state,

$$\mathbf{x}_0^g = \mathbf{x}^b. \tag{2}$$

For each loop $k \in [0,K]$, the non-linear cost function $\mathscr{J}(\mathbf{x}(t_0))$ is firstly linearised around the first guess states $\mathbf{x}_k^g$ leading to the linear cost function J_k, and secondly J_k is expressed in terms of the increment $\delta\mathbf{x}_k$ between the model state and the first guess,

$$\delta\mathbf{x}_k = \mathbf{x}(t_0) - \mathbf{x}_k^g. \tag{3}$$

If $\delta\mathbf{x}_k^a$ is the increment that minimises the linear cost function J_k for the loop k, i.e.

$$\delta\mathbf{x}_k^a = \arg\min_{\delta\mathbf{x}_k}\left[J_k(\delta\mathbf{x}_k)\right], \tag{4}$$

then the first guess state for the next loop $(k+1)$ is defined as

$$\mathbf{x}^g_{k+1} = \mathbf{x}^g_k + \delta\mathbf{x}^a_k \,. \tag{5}$$

2.1.2 *Linearisation of the cost function*

To use the increment $\delta\mathbf{x}_k$ of Eq. (3) for the linearisation of the cost function of Eq. (1), one has to first linearise the model equivalent of the observation at time t_i,

$$\begin{aligned} H_i \circ M_i\left(\mathbf{x}\left(t_0\right)\right) &= H_i \circ M_i\left(\mathbf{x}^g_k + \delta\mathbf{x}_k\right) \\ &= H_i \circ M_i\left(\mathbf{x}^g_k\right) + \mathbf{H}_i\mathbf{M}_i\,\delta\mathbf{x}_k + \mathscr{O}\left(\delta\mathbf{x}_k\right), \end{aligned} \tag{6}$$

where $\mathbf{H}_i$ and $\mathbf{M}_i$ are the tangent-linear version of H_i and M_i respectively, linearised around $\mathbf{x}^g_k$. Note that $\mathbf{H}_i$ and $\mathbf{M}_i$ are varying with each loop k and should be indexed with k. We nevertheless omit the index k to simplify the notation.

To further simplify the expression of the cost function, we also introduce the innovation vector for each outer loop k, i.e. the difference between observation and model equivalent from the first guess state of the current loop $\mathbf{x}^g_k$,

$$\mathbf{d}^{o,g}_{i,k} = \mathbf{y}^o_i - H_i \circ M_i\left(\mathbf{x}^g_k\right) . \tag{7}$$

When neglecting the second order terms $\mathscr{O}\left(\delta\mathbf{x}_k\right)$, the linearisation of the cost function of Eq. (1) around the first guess state $\mathbf{x}^g_k$ leads to

$$\begin{aligned} J_k\left(\delta\mathbf{x}_k\right) &= \frac{1}{2}\left(\mathbf{x}\left(t_0\right) - \mathbf{x}^b\right)^T \mathbf{B}^{-1}\left(\mathbf{x}\left(t_0\right) - \mathbf{x}^b\right) \\ &+ \frac{1}{2}\sum_{i=0}^{N}\left[\mathbf{H}_i\mathbf{M}_i\,\delta\mathbf{x}_k - \mathbf{d}^{o,g}_{i,k}\right]^T \mathbf{R}_i^{-1}\left[\mathbf{H}_i\mathbf{M}_i\,\delta\mathbf{x}_k - \mathbf{d}^{o,g}_{i,k}\right] , \end{aligned} \tag{8}$$

where $\mathbf{x}\left(t_0\right) - \mathbf{x}^b = \mathbf{x}^g_k - \mathbf{x}^b + \delta\mathbf{x}_k$.

The minimisation of a linear cost function J_k of Eq.(8) with an iterative process is known as inner loop. The iterative process by which the non-linear cost function $\mathscr{J}$ is linearised around a more accurate first guess state is known as outer loop.

2.1.3 *Change of variable*

To improve the condition number of the linear cost function and to accelerate the convergence, we introduce the same change of variable as already presented in Massart (2018), with the control vector χ,

$$\chi = \mathbf{B}^{-\frac{1}{2}}\left(\mathbf{x}\left(t_0\right) - \mathbf{x}^b\right) \quad \text{or} \quad \mathbf{x}\left(t_0\right) = \mathbf{x}^b + \mathbf{B}^{\frac{1}{2}}\chi \,. \tag{9}$$

Replacing $\mathbf{x}\left(t_0\right)$ in Eq. (3) using Eq. (9) leads to

$$\delta\mathbf{x}_k = \mathbf{B}^{\frac{1}{2}}\chi + \mathbf{x}^b - \mathbf{x}^g_k \,. \tag{10}$$

The linear cost function of Eq. (8) can then be expressed in terms of the control vector χ,

$$\begin{aligned} J_k\left(\chi\right) &= \frac{1}{2}\chi^T\chi \\ &+ \frac{1}{2}\sum_{i=0}^{N}\left[\mathbf{H}_i\mathbf{M}_i\left(\mathbf{B}^{\frac{1}{2}}\chi + \mathbf{x}^b - \mathbf{x}^g_k\right) - \mathbf{d}^{o,g}_{i,k}\right]^T \mathbf{R}_i^{-1}\left[\mathbf{H}_i\mathbf{M}_i\left(\mathbf{B}^{\frac{1}{2}}\chi + \mathbf{x}^b - \mathbf{x}^g_k\right) - \mathbf{d}^{o,g}_{i,k}\right] . \end{aligned} \tag{11}$$

Its gradient with respect to the control vector χ is

$$\nabla_{\chi} J_k(\chi) = \chi + \sum_{i=0}^{N} \mathbf{M}_i^T \mathbf{H}_i^T \mathbf{R}_i^{-1} \left[\mathbf{H}_i \mathbf{M}_i \left(\mathbf{B}^{\frac{1}{2}} \chi + \mathbf{x}^b - \mathbf{x}_k^g \right) - \mathbf{d}_{i,k}^{o,g} \right], \tag{12}$$

where $\mathbf{M}_i^T$ and $\mathbf{H}_i^T$ are the adjoint versions of $\mathbf{M}_i$ and $\mathbf{H}_i$ respectively.

2.1.4 *Incremental approach*

Another aspect of the incremental formulation is the possibility to use a lower resolution version of the tangent-linear operators $\mathbf{H}_i$ and $\mathbf{M}_i$ (and their adjoint) for the inner loops of the minimisation process (Fig. 1). It is usual to start with a resolution much lower than the one of the non-linear model and to increase the inner loop resolution with each outer loop. In that case, $\mathbf{x}^b$ and $\mathbf{x}^g$ are interpolated to the resolution of the inner loops and the product $\mathbf{B}^{\frac{1}{2}} \chi$ is also computed at the resolution of the inner loops.

In the current formulation, the control vector χ is truncated in the wavelet space at the spectral resolution of the inner loops. In order to express the product $\mathbf{B}^{\frac{1}{2}} \chi$ at the resolution of the inner loops, there is a specific treatment for $\mathbf{B}^{\frac{1}{2}}$ to ensure that it represents the covariances at that resolution.

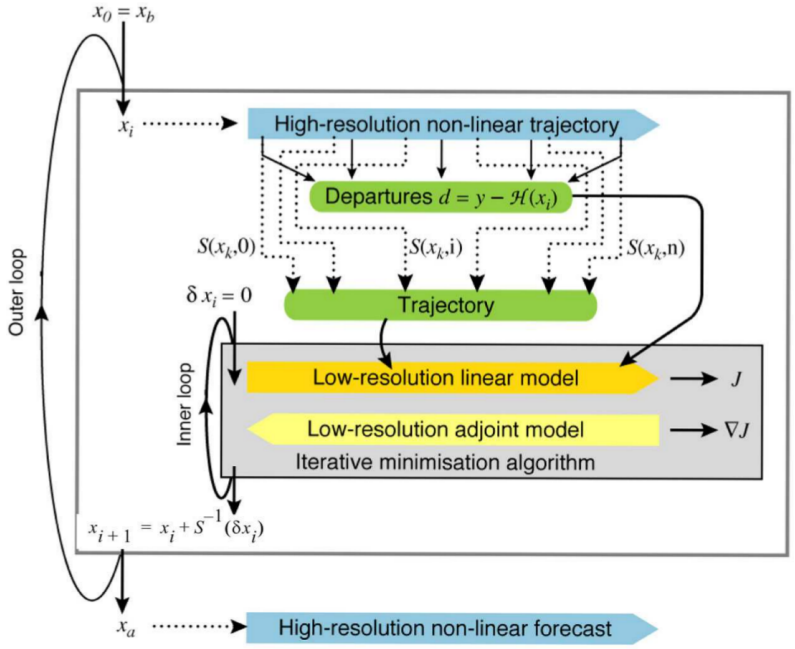

Figure 1: Schematic of the IFS 4D-Var incremental formulation algorithm. Outer loops are performed at high resolution using the full non-linear model. Inner iterations are performed at lower resolution using the tangent-linear forecast model, linearised around a 12-hour succession of model states (the trajectory) obtained through interpolation from high resolution (S denotes the interpolation operator, J the cost function and **x** *the atmospheric state vector). Figure extracted from the IFS documentation (https://www.ecmwf.int/node/19306, Chapter 2: 4D variational assimilation).*

2.2 Incremental formulation with the new hybrid background error model

2.2.1 Algorithm

The α-$\mathbf{B}^h$ formulation implemented in the IFS by Massart (2018) consists in a modification of the change of variable of Eq. (9) with

$$\mathbf{x}(t_0) - \mathbf{x}^b = \gamma^s \mathbf{B}^{s\frac{1}{2}} \chi^s + \gamma^e \mathbf{B}^{e\frac{1}{2}} \chi^e , \tag{13}$$

where χ^s and χ^e are the parts of the control vector χ that use the static background errors $\mathbf{B}^s$ and the ensemble-based background errors $\mathbf{B}^e$ respectively. The weights γ^s and γ^e have to satisfy the relationship $\gamma^{s2} + \gamma^{e2} = 1$.

Using Eq. (3) in Eq. (13) leads to an increment $\delta\mathbf{x}_k$ for the loop k of the form

$$\delta\mathbf{x}_k = \gamma^s \mathbf{B}^{s\frac{1}{2}} \chi^s + \gamma^e \mathbf{B}^{e\frac{1}{2}} \chi^e + \mathbf{x}^b - \mathbf{x}^g_k . \tag{14}$$

Comparing Eq. (14) to Eq. (10) shows that the incremental approach does not need any specific implementation for the α-$\mathbf{B}^h$ formulation in terms of algorithm. For the next outer loop $k+1$, the first guess is then a combination of the static and ensemble-based background errors,

$$\mathbf{x}^g_{k+1} = \mathbf{x}^b + \gamma^s \mathbf{B}^{s\frac{1}{2}} \chi^{a,s}_k + \gamma^e \mathbf{B}^{e\frac{1}{2}} \chi^{a,e}_k , \tag{15}$$

where $\chi^{a,s}_k$ and $\chi^{a,e}_k$ are the parts of the control vector χ^a_k corresponding to χ^s and χ^e respectively and resulting from the minimisation in the current loop k.

Just like in the current implementation, in case of change of inner loops resolution from one outer loop to the next, the product $\mathbf{B}^{e\frac{1}{2}} \chi^e$ has to be at the right resolution. First, χ^e is truncated in the wavelet space at the spectral resolution of the inner loops as it is done for the other parts of χ. Concerning $\mathbf{B}^e$, there are two aspects to consider: the resolution of the used EDA members and the localisation length-scale. These two aspects are treated separately in the next two sections.

2.2.2 Resolution of the perturbations

The α-$\mathbf{B}^h$ formulation makes a direct use of the M members $\mathbf{x}^b_m$, $\forall m \in [1, M]$ of the EDA. The mean value is removed from each member to build M perturbations. Currently all members of the EDA have the same spectral resolution of TCo 639, which corresponds roughly to a grid spacing of $\Delta x \approx 20\,\text{km}$. When using the α-$\mathbf{B}^h$ formulation together with the incremental formulation, the background error matrix $\mathbf{B}^e$ should be interpolated at the inner loop resolution. This means that the EDA members have to be interpolated and used at the resolution of the inner loop.

In the following, the interpolation is carried out by extracting the EDA members from the ECMWF's archive and specifying the spectral resolution and the grid. We realised afterwards that this interpolation is not equivalent to the spectral truncation used to build the current background errors at different resolutions (for which the spectral coefficients corresponding to the wavenumbers larger than the truncation are set to zero). To illustrate the impact of the interpolation used in this work on the background errors, we computed the background error variances σ^{b2}_j for a given resolution j,

$$\sigma^{b2}_j = \frac{1}{M-1} \sum_{m=1}^{M} \left(\mathbf{T}_j \mathbf{x}^b_m - \mu_j \right)^2 , \quad \text{with} \quad \mu_j = \frac{1}{M} \sum_{m=1}^{M} \mathbf{S}_j \mathbf{x}^b_m , \tag{16}$$

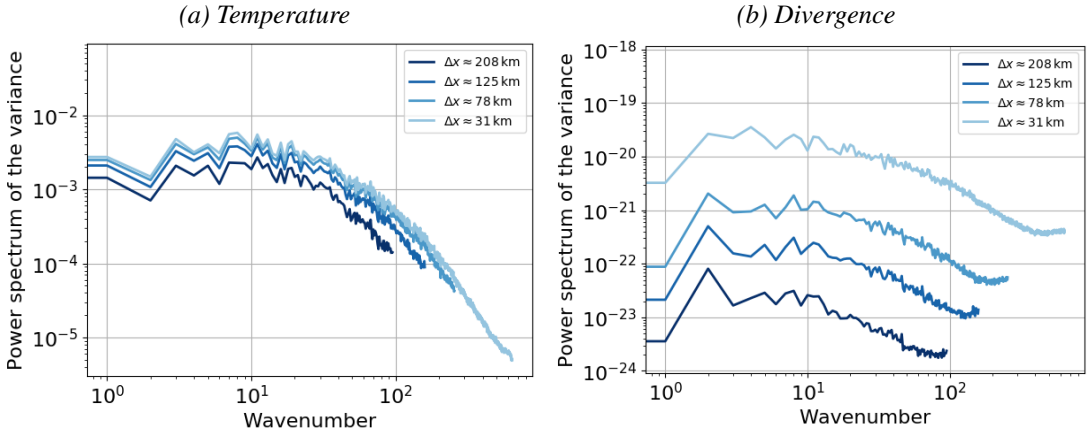

Figure 2: Examples of power spectrum of the estimated background error variance for (a) temperature and (b) divergence at model level 118 (around 900 hPa*). Each line corresponds to the variance computed from the EDA members truncated at a given spectral resolution (see label for the approximate spatial resolution of each spectral truncation).*

where $\mathbf{T}_j$ is the interpolation at the truncation of the wavenumber j.

Figure 2 presents the power spectrum of $\sigma_j^{b^2}$ for temperature and divergence at model level 118 ($\approx$ 900 hPa) for the interpolation $\mathbf{T}_j$ at various truncations ranging from an equivalent grid spacing of $\Delta x \approx 30$ km to $\Delta x \approx 200$ km. If a spectral truncation had been used, all the curves would have been on top of each other until the wavenumber of the truncation. With the current interpolation, for temperature, the background error variance for each wavenumber is only slightly affected by the interpolation. On the other hand, the power spectrum of the variance for divergence varies significantly with the truncation and the variations are happening at all wavenumbers.

We have tested the effect of having a spectral truncation instead of the truncation performed by the retrieving process from the archive. We found that the impact is relatively small compared to other signals (not shown).

2.2.3 Localisation length-scale

In order to reduce the sampling noise induced by using a limited number of EDA members in the new hybrid formulation, we documented in Massart (2018) the need for a localisation function. We also presented how to diagnose the optimal value of the localisation length-scale (under some assumptions), using the *HybridDiag* software developed by B. Ménétrier and presented in Ménétrier et Auligné (2015).

Here, we want to assess the sensitivity of the diagnosed optimal value of the localisation length-scale with respect to the spectral truncation of the EDA members. As we did for the variances in the previous section, we applied a given spectral truncation to the EDA members and we computed the corresponding localisation length-scale from these truncated members. We then repeated this operation for the set of spectral truncations used previously and we accordingly have a diagnosed value of the localisation length-scale per spectral truncation.

We found that for each spectral truncation, the diagnosed optimal value of the localisation length-scale varies little with altitude between 1000 hPa and 100 hPa (not shown). We therefore computed the mean

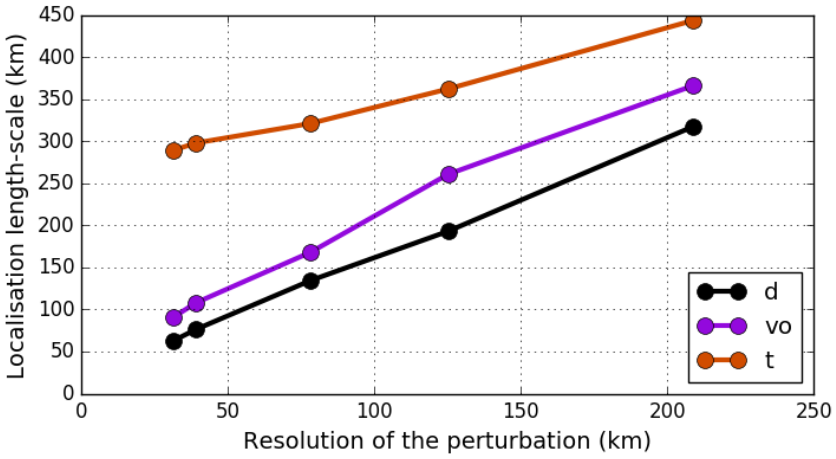

Figure 3: Impact of the resolution of the EDA members on the diagnosed mean localisation length-scale for divergence (black), vorticity (purple) and temperature (red). The mean localisation length-scale is the average of the diagnosed global localisation length-scale between 100 hPa *and 1000* hPa.

value along this pressure range. We found a sensitivity of the mean value to the spectral truncation for all the atmospheric model variables of the control vector (Fig. 3). The relationship is almost linear, the larger the truncation (i.e. the smaller the spatial resolution), the smaller the localisation length-scale as diagnosed by *HybridDiag*.

This result suggests that one should use a different value of the localisation length-scale for each outer loop when the resolution of the inner loops changes. For example, for an experiment with 3 outer loops and with the inner loops resolutions being TL95 / TL159 / TL255 (or spatial resolution of about 200 km, 125 km and 80 km respectively), according to *HybridDiag* one should use the following value for the localisation length-scale of temperature for each outer loop: 450 km, 350 km and 300 km.

3 Evaluation of the incremental formulation with a resolution-dependent localisation length-scale

In the previous section we found that the new α-$\mathbf{B}^h$ formulation for the background error covariances can be applied together with the standard incremental formulation. We also found that the localisation length-scale should be dependent of the inner loop resolution of each outer loop. In this section, we present the effect of a resolution-dependent localisation length-scale, starting with single-observation experiments and then using the full observing system.

3.1 Single observation experiment

We mimic here the single observation experiment HYB_100_75m of Massart (2018), where a single observation of temperature is assimilated at the beginning of the assimilation window. However, we changed the location of the observation to a region where the uncertainty on surface pressure (as derived from the EDA) is higher, i.e. $55°N, 68°W$, for the selected date (1 November 2016). The altitude of the observation is still 900 hPa.

The observation is assumed to be an aircraft observation. This means that the observation operator is linear. Moreover, the observation being located at the beginning of the assimilation window, the forecast model does not have any influence on the increment.

3.1.1 Experimental setup

We are using here the incremental formulation with three outer loops. For the successive outer loops, the inner loop resolutions are TL 95, TL 159 and TL 255. For this configuration, we have two experiments, both with an hybrid weight $\gamma^{e2} = 1$ (fully ensemble-based background errors). The only difference between the two experiments is the value of the localisation length-scale used for each outer loop. For the first experiment, referred to as *Cst. Loc.*, the localisation length-scale is constant with a value of 300 km. The second experiment, referred to as *Var. Loc.*, has varying localisation length-scale for each outer loop: from 800 km for the first outer loop to 480 km for the second and 300 km for the third. These values are chosen to be larger than the ones found in the previous section for the first two outer loops in order to better illustrate the possible effect of resolution-dependent localisation length-scale.

3.1.2 Results

Figure 4 displays the temperature increments for the first experiment after each of the three outer loops and the total increment at the model level closest to the observation level. The first outer loop produces a negative increment with a large-scale almost isotropic shape due to the broad resolution of the loop (TL 95). The second and third outer loops produce much smaller scale increments with positive and negative values about an order of magnitude less than the values of the increment from the first loop. The total increment is then similar to the one from the first loop but with a more refined structure due to the increased resolution and the refinement of the smaller structures from the second and third outer loops.

For the second experiment with varying localisation length-scale, the increment after the first loop has a larger scale structure than the one from the previous experiment (Figs. 4a and 5a). This is expected as the localisation length-scale is more than twice larger. The second and last loop have then much work to do by first removing the increment far (with respect to the localisation length-scale) from the observation location and secondly refining the smaller scale structures close to the observation location (Figs. 5b and 5c).

After the three outer loops, the total increment is nonetheless similar in shape and in amplitude in the two experiments (Figs. 4d and 5d). For this simple experiment with a single observation at the beginning of the window and a linear observation operator, only the localisation length-scale of the last loop matters. This statement cannot be generalised to the situation for which we are assimilating all the available observations. The model and the observation operators are indeed linearised around the first guess for the current outer loop. As the first guess is dependent on the localisation length-scales used within the previous outer loops, the successive linear versions of the non linear cost function are therefore also dependent on the localisation length-scales used within the previous outer loops.

3.2 Full observing system experiment

To better understand the implication of a resolution-dependent localisation length-scale, let us take the example of the two previous experiments: one with a constant localisation length-scale and one with a resolution-dependent one. For the two experiments, having a different localisation length-scale in the first

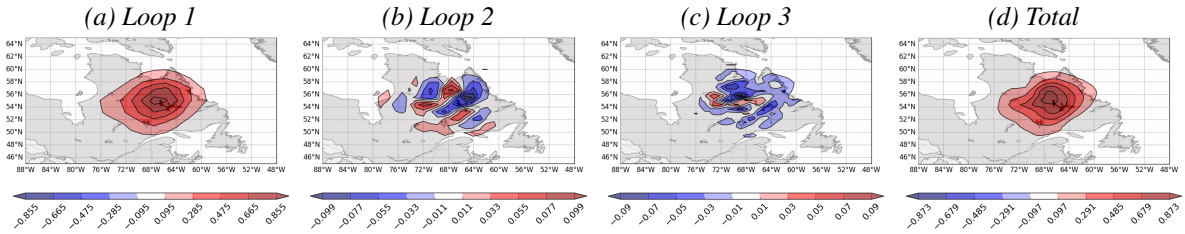

Figure 4: Example of temperature increments after each of the 3 outer loops (a to c) and total increment (d) obtained from a single observation of temperature localised in the middle of the domain and for the hybrid experiment with the same localisation length-scale for all the outer loops.

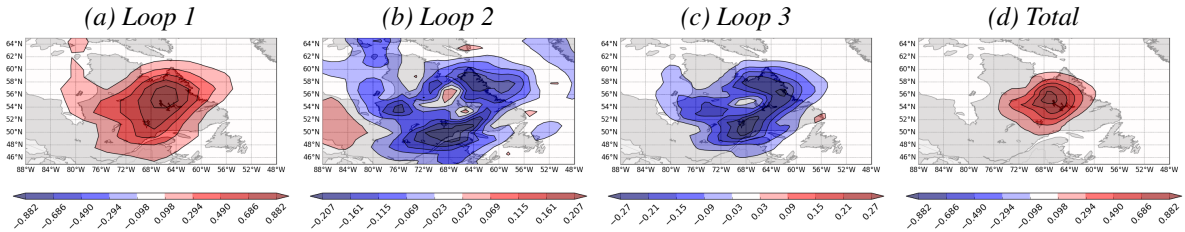

Figure 5: Same as Fig. 4 but with varying localisation length-scale for each outer loop: from 800 km for the first outer loop to 480 km for the second and 300 km for the third.

outer loop results in having a different analysis increment. For each experiment, its analysis increment is added to the background to build the first guess and the new model trajectory for the next outer loop. They are both unique to the experiment and therefore the tangent-linear and adjoint versions of the model and observation operators are also specific to the experiment. This implies that the linear cost function of Eq. (8) to minimise for the second loop could be considerably different between two experiments. The same logic applies for the next outer loops.

The non-linear cost function $\mathscr{J}$ is nonetheless the same for the two experiments. For a mildly non-linear cost function with one global minimum and an infinite number of outer loops, the analyses should be the same for both experiments and should be the global minimum of $\mathscr{J}$. The linear cost functions being different between the experiments for each outer loop, the path to reach the global minimum is also different. Therefore, for a finite number of outer loops, we may expect to find a different analysis that would correspond to a local minimum along the path determined by each experiment.

3.2.1 Experimental setup

To assess the impact of the localisation length-scale on the final analysis, we now assimilate all observations available on 1^{st} November 2017, 00 UTC. Using a full ensemble-based background error was found not to be optimal and a regularisation from a static background error is desirable. For that purpose, we use a weight γ^{e2} of 75%, which allows a significant part of increment to come from the ensemble-based part of the background error covariances, while still providing sufficient regularisation.

3.2.2 Results

To compare the experiments, we first examine the spectral density of the increments for all the atmospheric model variables and all model levels. For the 3D variables, we compute the difference in spectral space between the constant localisation (*Cst. Loc.*) and the variable localisation (*Var. Loc.*) experiments, as a function of model level.

For the total increment of the logarithm of surface pressure (Fig. 6), the localisation length-scale has an impact mainly on the high wavenumbers (small scale structures). The experiment with the largest localisation length-scale has more energy after the first outer loop. The difference further increases with the next outer loops with an offset in the spectra starting from wavenumber 20. For the low wavenumbers (large scale structures), there is a difference after the first outer loop, but the difference tends to shrink with the next outer loops, and the total increment is similar for the two experiments in this spectral region.

The behaviour of the increments for all the other atmospheric model variables and for all levels is similar to the one of the increment of the logarithm of surface pressure (Fig. 7):

1. low wavenumbers: no clear pattern for the difference between the two experiments (for all model levels and all model variables) and the difference is shrinking with the number of outer loop ;
2. high wavenumbers: positive offset in the spectral density of the increments for the experiment with the variable localisation length-scale starting from wavenumber 20 and offset increasing with the number of outer loop.

To understand how these differences in the increments impact the analysis, we can examine some diagnostics in observation space. First, we compute the analysis fit to observation (i.e. "observation" part of the cost function) normalised by the initial value of the cost function, as a function of the outer loop index (Fig. 8a). Note that we have two values for the second and third outer loop: one value after adding the analysis increment (first guess) and one value after minimisation (analysis).

After the first loop, the experiment with the constant localisation length-scale slightly outperforms the experiment with the variable localisation length-scale. The difference gets less and less with the next outer loops. This means that the localisation length-scale does not have a strong impact on how the system fits the observations globally. When looking more in detail, the experiment with the varying localisation length-scale has a better fit to almost all observations except the radar rain rates and all-sky satellite radiances (not shown). This could be linked to the fact that the observation operators are non-linear for these data and the experiment with a resolution-dependent localisation length-scale provides lower quality first-guesses fields in first outer loop due to a broader length-scale.

The second diagnostic is the reduction of the gradient norm as a function of the inner loop iteration, for the 3 outer loops (Figs 8b to 8d). The experiment with the constant localisation length-scale converges sightly faster than the one with the variable localisation length-scale for the first outer loop. The behaviour is then similar for the last two outer loops.

Overall, varying the localisation length-scale together with the resolution of the inner loops has an impact on the analysis, but the experimentation so far does not show if the impact is positive or not. This will be assessed again in future configurations of the α-$\mathbf{B}^h$ formulation. In the following, the localisation length-scale will be kept constant for two main reasons. First, the diagnosed localisation length-scale shows some clear regional variations. The priority would be to first use a space-dependent localisation length-scale before varying it with the inner loop resolution. Secondly, when the localisation length-scale varies between the outer loops, there is more energy in the small scale for all variables. The new

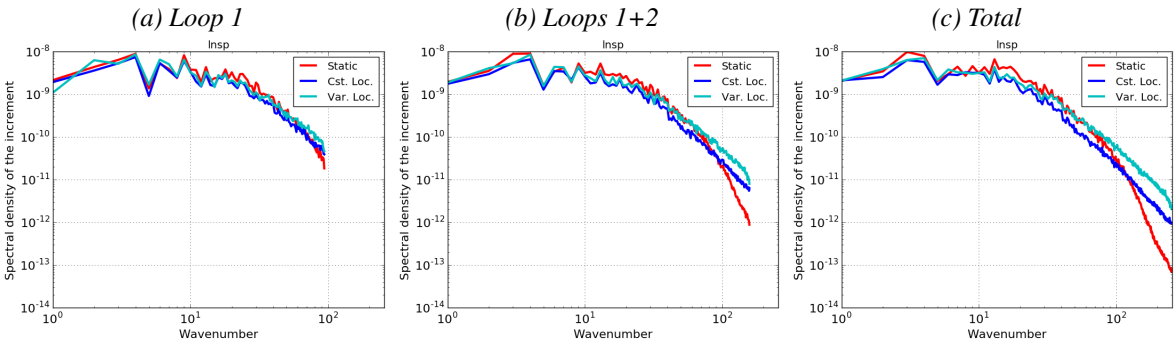

Figure 6: Spectral density of the total increment of the logarithm of surface pressure after each outer loop (a) to (c), for the static experiment (red) and the hybrid experiments with constant (blue) and varying (cyan) localisation length-scale.

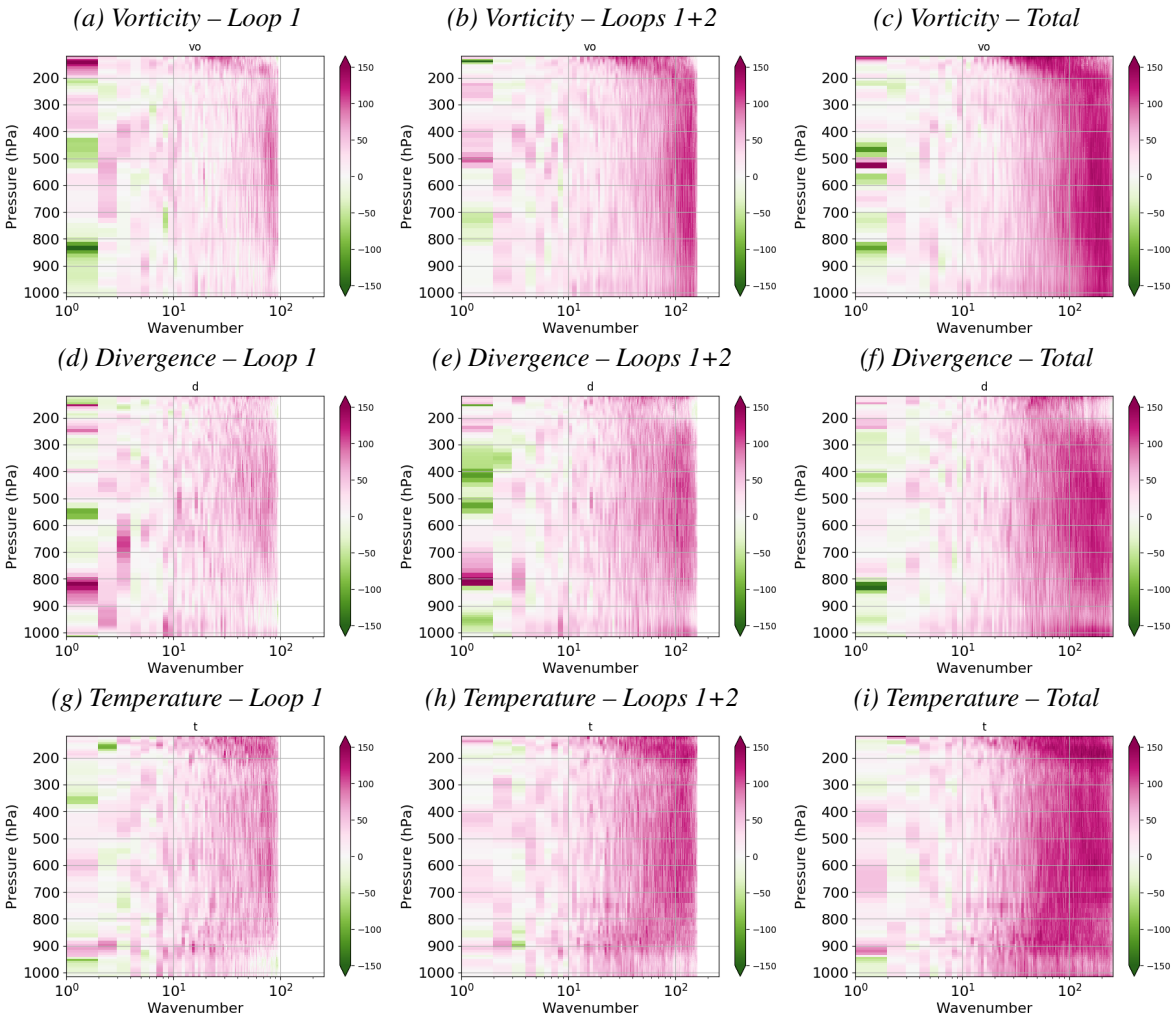

Figure 7: Difference (in %) in the spectral density of the total increment of the atmospheric model variables between the experiment with the varying localisation length-scale and the experiment with the constant localisation length-scale. From top to bottom: vorticity, divergence and temperature. From left to right: difference of the total increment after each outer loop from 1 to 3.

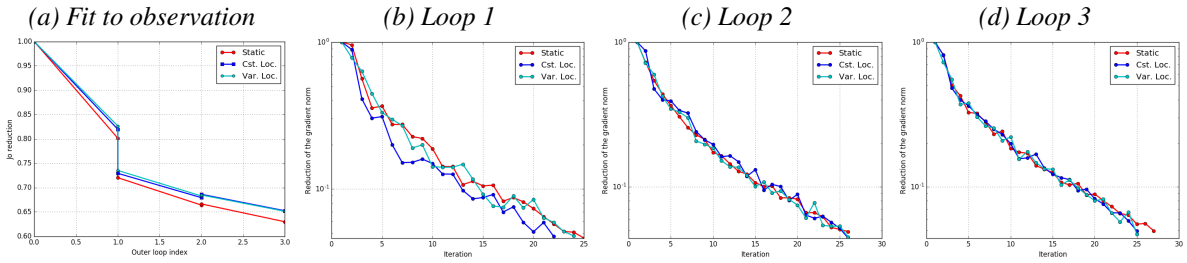

Figure 8: Impact of the localisation length-scale variation on the convergence of the 4D-Var: (a) reduction of the observation cost function as a function of the outer loop and (b) to (d): reduction of the gradient norm as a function of the inner loop iteration for each outer loop. In red: static experiment, in blue: hybrid experiment with fixed localisation length-scale (300 km), and in cyan: hybrid experiment with varying localisation length-scale (800 km/480 km/300 km). The hybrid experiments have a hybrid weight $\gamma^{e2} = 0.75$ here.

formulation already produces increments with more energy in the small scales when compared with the experiment with the static **B** matrix as illustrated for the logarithm of surface pressure (Fig. 6). It is believed that too much energy in the small scales may have a negative impact on the analysis. Based on this evidence, we choose to keep the localisation length-scale constant.

4 Evaluation of the new formulation over a winter season

The previous results with the single observation or the full observing system plus some additional preliminary tests over longer periods gave us confidence in the correctness of the implementation of the α-$\mathbf{B}^h$ formulation. This section assesses the benefits of the new approach compared to the current formulation by cycling the assimilation experiments over a period of 4 months.

4.1 Configuration

4.1.1 Experiments

Three experiments are discussed and compared in this section. The first experiment referred to as $\mathbf{B}^o$ uses the wavelet background errors as used in operation for CY45R1. The second experiment referred to as $\mathbf{B}^s$ uses static wavelet background errors (see section 4.1.2). The last one referred to as α-$\mathbf{B}^h$ uses the new hybrid formulation. We do not present here the sensitivity of the α-$\mathbf{B}^h$ formulation with, for example, the hybrid weight or the number of perturbations. We wanted to have a first assessment of the α-$\mathbf{B}^h$ formulation performance before trying the optimise its performance with more specific tuning.

All the three experiments share the same configuration except for the background error formulation. They are based on IFS CY45R1. The outer loop resolution is TCo 399 and the inner loop resolutions are successively TL 95 / TL 159 / TL 255, all using 137 vertical levels. The period during which they are compared spans from beginning of November 2017 to end of February 2018 (4 months of the Northern Hemisphere winter).

4.1.2 *Static background errors*

The α-$\mathbf{B}^h$ and $\mathbf{B}^o$ experiments have consistent background errors as they are both based on the perturbed members of same EDA. In order to have a consistent static background error covariance matrix, we computed it as an average of $\mathbf{B}^o$ matrix over the period of comparison (see Appendix A for more details). This means that on average the background error standard deviation of each of the atmospheric model variables is the same for all experiments over the period of interest. Nonetheless, the correlation length-scales are different between the three background error covariance matrices and we can expect sightly different effective covariances between the experiments.

Our strategy of building a static background error covariance matrix allows a fair comparison. This is nevertheless not applicable in an operational environment as the future background errors are unknown. In that particular situation, one could use the same method as the one used to building the climatological background error correlations and based on an average over about a year.

4.1.3 *Hybrid α-$\mathbf{B}^h$ background errors*

One main difference for the α-$\mathbf{B}^h$ formulation here compared to previous experiments is the use of time lag for the perturbations. Previously, we used 25 perturbations at the analysis time and ± 3 hour before (or the equivalent of $25 \times 3 = 75$ members). We now use hourly perturbations in the same time range of ± 3 hour (or the equivalent of $25 \times 7 = 175$ members).

(a) Diagnosed pressure of the tropopause *(b) Tapering zone*

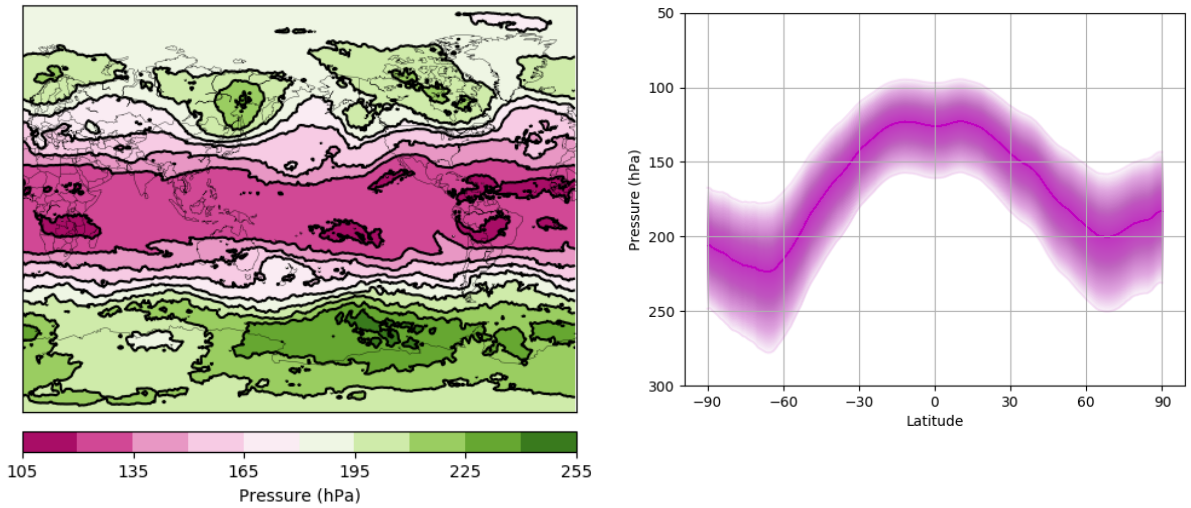

Figure 9: (a) Mean pressure level of the diagnosed "humidity" tropopause for November 2017. (b zonal and temporal mean of the zone where the tapering of the hybrid weight happens (shaded area). The hybrid weight is 0 above the shaded area, and 0.5 below. Within the shaded area, the weight varies according to Eq. (4.10) of Gaspari et Cohn (1999) where the logarithm of the pressure is used as a distance.

The current implementation of the α-$\mathbf{B}^h$ formulation impose to have the same horizontal localisation for all model levels. In Massart (2018), we showed that the global mean of the localisation length-scale as diagnosed by the *HybridDiag* software is almost constant in the troposphere and then varies much more with the vertical in the stratosphere. Therefore, it was decided not to apply the α-$\mathbf{B}^h$ formulation in the stratosphere. We first diagnose for each grid point the model level corresponding to the "humidity"

tropopause in the background field (model level for which the specific humidity is larger than 3 mg.kg^{-1} and the specific humidity two model levels below is larger than 5 mg.kg^{-1}, when looking for model levels located in the pressure range between 500 hPa and 70 hPa). Then, for each horizontal grid point, we are using a hybrid weight of 0.5 in the vertical up to 5 model levels below the "humidity" tropopause. Then the tapering starts and reaches zero at 5 model levels above the "humidity" tropopause (Fig. 9). This ensures that the formulation is used mainly in the troposphere, but relies on the ability to diagnose properly the "humidity" tropopause.

We perform the horizontal localisation on the EDA perturbations after applying an operator based on the variations of the Laplacian operator proposed by Berre *et al.* (2017). Unless specified otherwise, we are using the same transformations as the ones presented in in Massart (2018).

4.1.4 *Localisation of specific humidity perturbations*

The first results obtained with the α-$\mathbf{B}^h$ formulation showed a large increase of RMS error in relative humidity compared to the experiment with the static $\mathbf{B}$ matrix over the polar regions in the analysis and therefore in the first hours of the forecast (Fig. 10a). This was confirmed when the forecast from the two experiments (α-$\mathbf{B}^h$ and $\mathbf{B}^s$) were compared to relative humidity observations in the Arctic (Fig. 10b). The RMS error was up to 2% higher for the α-$\mathbf{B}^h$ experiment at 700 hPa and decreased until day 3 from which the change in RMS error was not significant anymore. For this reason, we restarted the α-$\mathbf{B}^h$ experiment and we localised the ensemble perturbations of specific humidity without using any transformation based on the Laplacian operator. This allowed to remove the observed increase in the RMS error in relative humidity. This is the version of the α-$\mathbf{B}^h$ experiment that is discussed in the rest of the document.

(a) Versus own analysis (T+12) *(b) Versus observations over Arctic (700 hPa)*

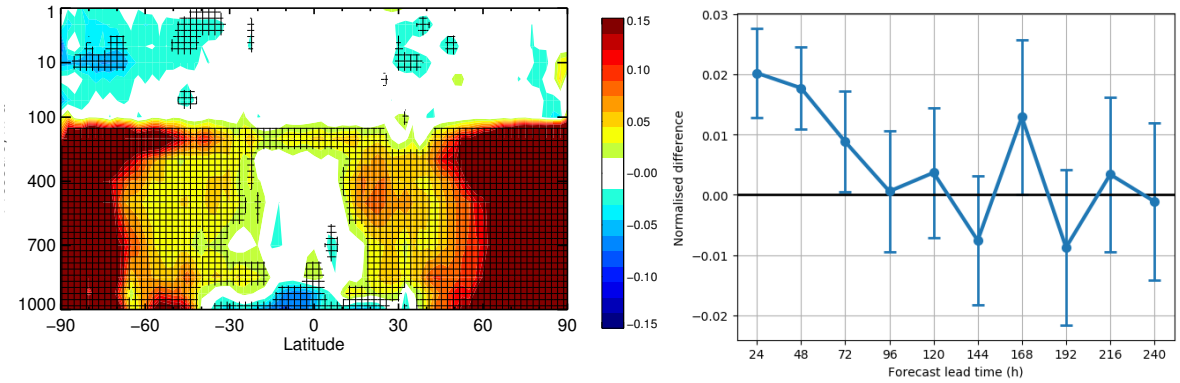

Figure 10: Change in RMS error in relative humidity between the α-$\mathbf{B}^h$ and the $\mathbf{B}^s$ experiments when the Laplacian operator is used for the localisation of the specific humidity in the α-$\mathbf{B}^h$ experiment. Positive values (or reddish colours) mean that the α-$\mathbf{B}^h$ experiment is worse than the $\mathbf{B}^s$ experiment.

4.2 Results

4.2.1 Convergence diagnostics

We first compare the convergence behaviour of the experiments. On average over the whole period of interest, the α-$\mathbf{B}^h$ experiment is the one that requires the least inner loop iterations to converge for each outer loop (Tab. 1). The $\mathbf{B}^s$ experiment requires between 1 and 2 more iterations, while the $\mathbf{B}^o$ experiment requires between 5 and 7 more iterations.

These convergence behaviours are consistent with the condition number diagnostics (Tab. 2). The α-$\mathbf{B}^h$ experiment has the lowest mean and median values of the condition number. Even for the last outer loop, there is an increase of almost 100% between the mean value of the condition number for the α-$\mathbf{B}^h$ experiment and the one for the $\mathbf{B}^o$ experiment, and an increase of 25% for the $\mathbf{B}^s$ experiment with respect to the α-$\mathbf{B}^h$ experiment.

We saw previously that the α-$\mathbf{B}^h$-like experiment produces increments with more energy in the small-scales and this was expected to negatively influence the convergence. Here, the α-$\mathbf{B}^h$ experiment still produces increments with more energy in the small-scales (Appendix B, page 36), yet its convergence is better. The energy in the small scales does not seem to perturb the convergence here.

The α-$\mathbf{B}^h$ experiment also had cases where the convergence was more difficult, as for the two other experiments. The condition number indeed reaches values much larger than its median value for some assimilation cycles. This is usually associated with increased number of inner loop iterations. There are no significant differences though between the α-$\mathbf{B}^h$ experiment and the two other experiments in terms of number of cycles with bad convergence over the period of interest.

To further investigate why the α-$\mathbf{B}^h$ experiment converges faster and has a lower condition number than the two other experiments, we consider now the reduction of the observation part of the cost function (or J^o) computed as the ratio between the value of J^o_k at the end of each outer loop k (using the analysis) divided by the initial J^o_0 value for the first outer loop (using the background). The J^o reduction is a global measure of how the analysis fits the observations.

The value of J^o and therefore the J^o reduction depends on the number of 'active' observations. Comparing the J^o reduction between experiments is significant only if the number of observations used by each experiment is similar. This is the case between the α-$\mathbf{B}^h$ experiment and the $\mathbf{B}^s$ experiment with an increase in the usage of observations by about 0.04% in the first experiment (Tab. 3). On the other hand, the $\mathbf{B}^o$ experiment uses about 1.2% less observations than the $\mathbf{B}^s$ experiment. This is not a large difference, but one has to be more cautious when comparing the J^o reduction of the $\mathbf{B}^o$ experiment against the two others.

As the mean value of the J^o reduction is similar to the median value over the period of interest, we just present the mean value (Tab. 4). The best fit-to-observation is for the analysis from the $\mathbf{B}^s$ experiment and the worst is from the $\mathbf{B}^o$ experiment. The α-$\mathbf{B}^h$ experiment is in-between and gets closer to the $\mathbf{B}^o$ experiment for the last loop. This means that on average over all observations and after 3 outer loops, the α-$\mathbf{B}^h$ experiment seems to perform similarly to the $\mathbf{B}^o$ experiment, but in less iterations. At the same time, the $\mathbf{B}^s$ experiment provides an analysis closer to the observations in just about one iteration more than for the α-$\mathbf{B}^h$ experiment. Having flow dependent background errors leads to slightly higher values for the analysis fit-to-observation.

Even if the global reduction of J^o is similar for the $\mathbf{B}^o$ and α-$\mathbf{B}^h$ experiments after 3 outer loops, they differ if we detail the diagnostic by observation type (Tab. 5). The α-$\mathbf{B}^h$ experiment has a worse fit-

Table 1: Basic statistics of the number of inner loop iterations for each outer loop minimisation and for the three experiments. In purple, the maximum value across the experiments for each statistic and in green the minimum value.

		$\mathbf{B}^s$	α-$\mathbf{B}^h$	$\mathbf{B}^o$
Min. 1	min.	23.0	**21.0**	**25.0**
	max.	31.0	**28.0**	**35.0**
	mean	26.3	**24.1**	**31.2**
	median	26.0	**24.0**	**31.0**
Min. 2	min.	24.0	**23.0**	**27.0**
	max.	28.0	**27.0**	**33.0**
	mean	26.1	**25.3**	**30.2**
	median	26.0	**25.0**	**30.0**
Min. 3	min.	26.0	**25.0**	**27.0**
	max.	38.0	**38.0**	**41.0**
	mean	27.5	**26.4**	**30.9**
	median	27.0	**26.0**	**31.0**

Table 2: Same as Tab. 1 but for the condition number.

$\mathbf{B}^s$	α-$\mathbf{B}^h$	$\mathbf{B}^o$
3 761.427	**2 195.577**	**8 575.633**
48 014.949	**88 298.863**	**38 273.039**
4 536.362	**3 445.418**	**13 266.499**
4 108.099	**2 615.873**	**12 184.716**
1 213.523	**1 180.399**	**2 958.827**
177 687 683.999	**5 810.697**	474 565.251
741 994.627	**1 379.798**	6 150.357
1 509.259	**1 310.957**	**3 902.236**
1 186.404	**1 160.257**	**1 990.397**
206 490.115	**128 341.354**	**252 694.669**
2 524.490	**2 019.636**	**3 885.214**
1 510.029	**1 261.863**	**2 679.927**

Table 3: Total number of 'active' observations for each outer loop (in %) normalised by the total number from the $\mathbf{B}^s$ experiment. In purple, the maximum value across the experiments, and in green the minimum value.

Minimisation	α-$\mathbf{B}^h$	$\mathbf{B}^o$
1	**100.040**	**98.350**
2	**100.041**	**98.762**
3	**100.041**	**98.763**

Table 4: Mean value of the reduction of J^o after each outer loop minimisation with respect to the initial J^o and for the three experiments. In purple, the maximum value across the experiments, and in green the minimum value.

Minimisation	$\mathbf{B}^s$	α-$\mathbf{B}^h$	$\mathbf{B}^o$
1	**0.803**	0.807	**0.809**
2	**0.665**	0.669	**0.671**
3	**0.624**	0.631	**0.631**

Table 5: Mean of the reduction of J_o by observation type for the last minimisation with respect to the initial J_o and for the three experiments. In purple, the maximum value across the experiments, and in green the minimum value.

Observation type	$\mathbf{B}^s$	α-$\mathbf{B}^h$	$\mathbf{B}^o$
Land stations and ships	**0.474**	0.479	**0.482**
Aircraft data	**0.502**	0.511	**0.513**
Atmospheric motion winds	**0.674**	**0.693**	0.678
Drifting Buoys	**0.366**	0.371	**0.374**
Radiosondes	**0.400**	**0.415**	0.409
Balloons and profilers	**0.364**	**0.375**	0.374
Satellite sounding data	**0.748**	0.750	**0.752**
Scatterometer	**0.509**	**0.528**	0.516
Limb observations	**0.563**	0.573	**0.576**
Ground-based precip. composites	**0.366**	**0.390**	0.385
All-sky satellite radiances	**0.499**	**0.511**	0.510

Table 6: Total number of 'active' observations used in Tab. 5 (in %) normalised by the total number from the $\mathbf{B}^s$ experiment.

α-$\mathbf{B}^h$	$\mathbf{B}^o$
100.00	**100.00**
100.00	100.00
100.01	**100.07**
100.00	**100.00**
100.00	**100.03**
100.00	**100.01**
100.05	**98.43**
100.00	**100.00**
100.00	**100.00**
100.34	**100.35**
100.01	100.01

to-observation for the data types related to wind measurements (atmospheric motion winds and scatterometer) and related to in-situ measurements (radiosondes, balloons and profilers). Otherwise, the $\mathbf{B}^s$ experiment has a systematic better fit-to-observation than the two other experiments for all observations.

The number of 'active' observations is similar in all experiments (within $\pm 0.3\%$) for all observation types except for the satellite sounding data (Tab. 6). For these particular data, the $\mathbf{B}^o$ experiment has on average 1.5% less 'active' data. Most of the difference comes from AMSU-A for which the number of 'used' data is similar between the experiments (not shown). The α-$\mathbf{B}^h$ and the $\mathbf{B}^s$ experiments have then more 'active' data due to a different first-guess check (based on the background error standard deviation) but the variational quality control gives little weight to these additional data.

4.2.2 Analysis

In this sub-section we investigate in more details the analysis departure. The mean analysis departure can be interpreted as the bias of the analysis compared to the observations. The analysis departure provides a limited information on the analysis as it is computed in observation space. The direct comparison between analyses from the different experiments provides a more comprehensive picture of the impact of the background error choice on the analysis. We are hence using the $\mathbf{B}^s$ experiment as the reference to compute the analysis difference with the two other experiments. Then, we compute the mean value and the standard deviation of the analysis differences.

We saw previously that the increments from an α-$\mathbf{B}^h$-like experiment have more energy in the small scales than those of a $\mathbf{B}^s$-like experiment after one assimilation cycle. To assess if this remains true in the context of the current experiments, we computed the mean spectra of the analysis fields at 0Z and 12Z for each of the three experiments. We then compute the difference against the mean spectra from the $\mathbf{B}^s$ experiment and normalise it by the mean spectra of the $\mathbf{B}^s$ experiment. We present in Appendix B.2 (page 37) the normalised spectra for the atmospheric model variables for the month of November only, as an average over a longer period could include a seasonal signal. The general structure of the normalised spectra are, however, similar for the other months.

The analyses at 0Z and 12Z are fields at the spectral truncation of 399 resulting from a short range forecast of 3 hour based on an increment at the spectral truncation of 255. This means that for the wavenumbers lower than 255, the analysis is directly constrained by the increment, and for the wavenumbers greater than 255, the analysis is the result of how the model absorbs the increment in the short range forecast.

In the next sections, we compare the analyses from the three experiments both directly in the physical or spectral space, or indirectly using the analysis departure. We first start with the model variables linked to the dynamics and then with the ones linked to the thermodynamics.

Dynamics

To better understand why the analysis from α-$\mathbf{B}^h$ experiment has a worse fit-to-observation in terms of J^o reduction than the two other experiments when dealing with observations related to wind measurements, we first investigate the analysis departure with the ASCAT and AMVs data. ASCAT provides data that can be compared to the model u and v components of wind at the altitude of 10 m. The Atmospheric Motion Vectors (AMVs) provide a picture of the wind above the surface.

Globally, the α-$\mathbf{B}^h$ experiment has the lowest mean value of the ASCAT analysis departure except for the northern hemisphere (Tab. 7). Elsewhere, the $\mathbf{B}^o$ experiment has the highest mean value of the ASCAT

Table 7: Mean value and standard deviation of the analysis departure (in $\mathrm{m.s^{-1}}$*) with respect to ASCAT observations per region and per analysis experiment. In purple the maximum value of the statistic per region, and in green the minimum value. The last column is the observation partition per region (in %).*

	Mean			Std. dev.			Obs.
Region	$\mathbf{B}^s$	α-$\mathbf{B}^h$	$\mathbf{B}^o$	$\mathbf{B}^s$	α-$\mathbf{B}^h$	$\mathbf{B}^o$	part.
S. Hemis.	-0.0623	**-0.0566**	**-0.0645**	**0.8037**	**0.8088**	0.8055	43.63
tropics	-0.0323	**-0.0281**	**-0.0568**	**0.8155**	**0.8339**	0.8179	31.50
N. Hemis.	-0.2082	**-0.2147**	**-0.2075**	**0.9249**	**0.9312**	0.9311	24.87
Global	-0.0891	**-0.0869**	**-0.0977**	**0.8412**	**0.8519**	0.8447	100.00

analysis departure. The main difference is found in the tropics where the α-$\mathbf{B}^h$ experiment has a bias of almost half of the one from the $\mathbf{B}^o$ experiment. This is due to a stronger (negative) u component of wind in the α-$\mathbf{B}^h$ experiment, when the $\mathbf{B}^o$ experiment has overall a weaker u component of wind and a weaker (positive) v component of wind (Fig. B.1, page 36). This signal is particularly strong west of the Andes (not shown).

The α-$\mathbf{B}^h$ analysis is the closest to the ASCAT observations on average, but at the same time the analysis departure from this experiment has more spatial and temporal variability than the two other analyses. When compared to the $\mathbf{B}^s$ experiment, the difference in the standard deviation of the analysis departure with ASCAT does not show any particular pattern (not shown) and the increase is similar in all regions of the globe.

Just above the surface, the mean analysis departure with AMVs is similar for all experiments except in the tropics. There, the α-$\mathbf{B}^h$ analysis has slightly more bias than the two other experiments in the vertical region between 700 and 500 hPa, with an increase up to $0.05\,\mathrm{m.s^{-1}}$ for a bias around $0.3\,\mathrm{m.s^{-1}}$ (not shown). The α-$\mathbf{B}^h$ experiment has weaker (negative) u component of wind than the $\mathbf{B}^s$ experiment which is already too weak compared to the observations (Fig. B.1, page 36). The $\mathbf{B}^o$ experiment performs slightly better than the α-$\mathbf{B}^h$ experiment thanks to a stronger u component of wind around 850 hPa.

The standard deviation of the analysis departure with AMVs is globally higher by about 1% for the α-$\mathbf{B}^h$ analysis compared to the $\mathbf{B}^s$ analysis and sightly higher than the $\mathbf{B}^o$ analysis (Fig. C.1, page 38). The largest difference is found in the tropics.

Meanwhile, the spatial standard deviation of the increment of the u and v components of wind (not shown) and the spatial standard deviation of the u and v analysis (Figs. 11a and 11b) is smaller in the α-$\mathbf{B}^h$ experiment than in the $\mathbf{B}^s$ one. The change is significant mainly in the tropics and below 200 hPa.

These results are all consistent with a smaller background error covariance in the α-$\mathbf{B}^h$ experiment. Even if the three experiments have consistent background error variances as detailed in section 4.1.2, the effective covariance may vary due to the correlations. The localisation process in the α-$\mathbf{B}^h$ experiment may reduce the correlation length-scales and therefore lower the effective covariance when compared to the two other experiments. In contrast, the $\mathbf{B}^o$ experiment may have a more variability in the effective covariance when compared to the $\mathbf{B}^s$ experiment because of the flow-dependent correlations. This would be consistent with the increase in the spatial standard deviation in the troposphere when compared to the $\mathbf{B}^s$ experiment (Figs. 11c and 11d).

The spectral distribution of the energy of the α-$\mathbf{B}^h$ analysis compared to the $\mathbf{B}^s$ analysis shows less energy in the medium and large scales, especially for divergence (Fig. B.3a, page 37). Our interpretation is that having a lower effective covariance prevents the observations to bring as much information in the

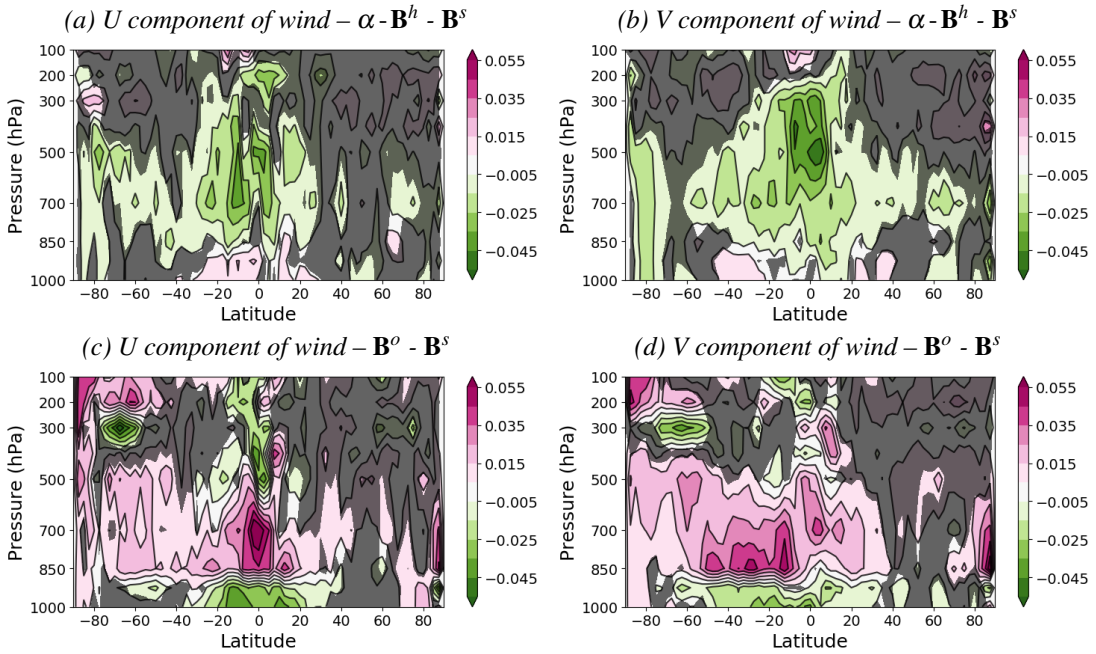

Figure 11: Change in standard deviation field in u (left) and v (right) component of wind (in $\mathrm{m.s}^{-1}$*) at initial time between the* α-$\mathbf{B}^h$ *and the* $\mathbf{B}^s$ *experiments (top) and between the* $\mathbf{B}^o$ *and the* $\mathbf{B}^s$ *experiments (bottom). The grey areas indicate a non statistically significant change.*

medium and large scales as a larger effective covariance would. A lower effective covariance restrains the analysis to fit too closely to the observations and then has more difficulty to reproduce the spatial variability that could be present in the observations. On the other hand, the α-$\mathbf{B}^h$ analysis has more energy in the small scales, starting from wavenumber around 100. To better understand this, a separate study on the modes of the background and observation error covariance matrices should be conducted. In the meantime, we recognise that overall the α-$\mathbf{B}^h$ analysis is smoother than the analyses of the two other experiments.

Thermodynamics

On average over the period of interest, there is little change in the mean fields in the temperature analysis between the α-$\mathbf{B}^h$ and the $\mathbf{B}^s$ experiments. The analysis from the α-$\mathbf{B}^h$ experiment is cooler by a few hundredths of a degree in the region around the tropopause, and also in the troposphere at mid-latitudes and in the tropics (Fig. B.2a, page 36). The largest differences occur in the Antarctic where the cooling around the tropopause reaches 0.09 K. When compared to radiosondes data, this cooling helps reducing the bias in the analysis (not shown). Where the analysis from the α-$\mathbf{B}^h$ experiment is cooler in the tropopause region, it is also wetter, but by not much more than 1%. The opposite is happening in the tropics where the α-$\mathbf{B}^h$ analysis is both cooler and drier in the mid-troposphere. Close to the surface, the α-$\mathbf{B}^h$ analysis becomes wetter but it is still cooler. When compared to radiosondes data, this cooling increases the analysis bias, while the increase in humidity reduces the bias.

The mean behaviour of the $\mathbf{B}^o$ experiment is similar to the one of the α-$\mathbf{B}^h$ experiment. The change in mean is nonetheless more pronounced for the relative humidity in the tropopause region where the the $\mathbf{B}^o$ analysis is wetter by more than 2 % than the two others (Fig. B.2d, page 36). The comparison with

specific humidity measurements from radiosondes data suggests that this change in the $\mathbf{B}^o$ experiment makes its analysis more biased than the analyses from the two other experiments. There is also a small increase of temperature in the region between 500 hPa and 300 hPa which is not present in the α-$\mathbf{B}^h$ experiment, which seems to increase the analysis bias when compared to aircraft data.

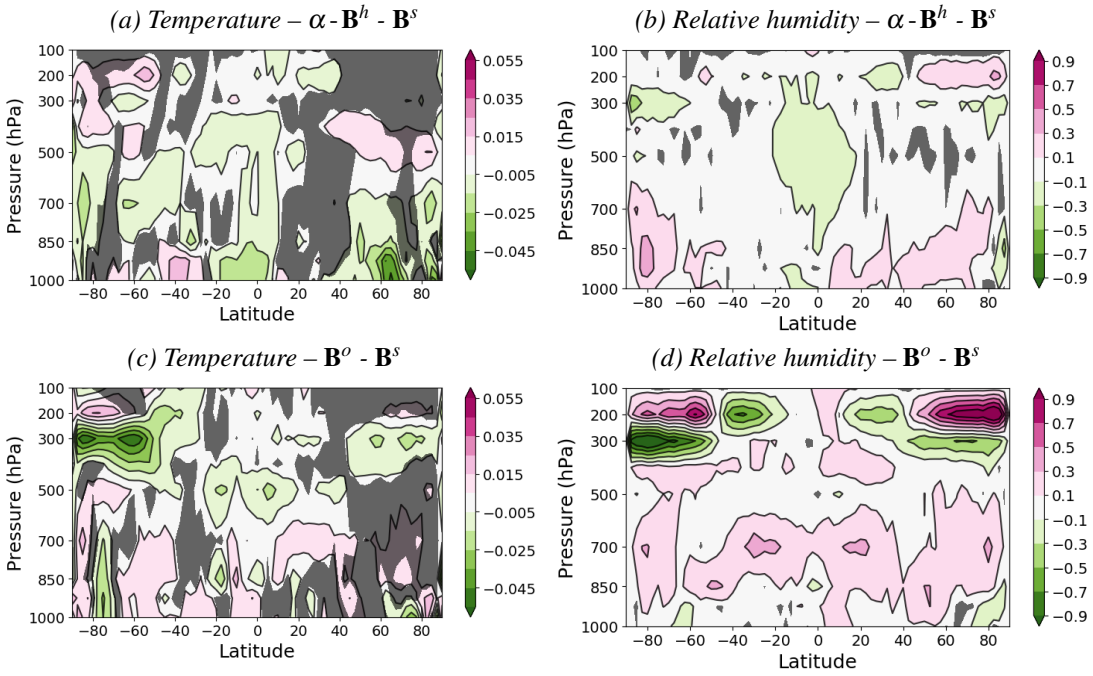

Figure 12: Change in standard deviation field in temperature (left, in k*) and relative humidity (right, in %) at initial time between the* α-$\mathbf{B}^h$ *and the* $\mathbf{B}^s$ *experiments (top) and between the* $\mathbf{B}^o$ *and the* $\mathbf{B}^s$ *experiments (bottom). The grey areas indicate a non statistically significant change.*

The standard deviations of the temperature analyses are very close to each other for the three experiments with differences of the order of 0.01K (Figs. 12a and 12c). The main significant difference comes from the $\mathbf{B}^o$ experiment where the variability is lower in the Antarctic region around 300 hPa by about 0.05K. This is associated with a lower variability of relative humidity by about 1 % (Fig. 12d). The small values of these differences explain why the mean J^o reduction is similar between the experiments for satellite data (Tab. 5).

4.2.3 First-guess

The forecast from the analysis is used in the next assimilation cycle to produce the background or initial first-guess state. The comparison of this first-guess with observations provides insight into the quality of the short range forecast. The observation part of the initial cost function at the beginning of each assimilation cycle naturally provides this comparison, accounting for the observation errors.

The first value of the initial cost function may vary a lot between assimilation cycles. This makes inappropriate the computation of statistics based directly on the value of the initial cost function. Instead, we choose the initial cost function of the $\mathbf{B}^s$ experiment after screening as a reference. The screening process is based on the first-guess state and therefore the number of used observations may vary between the experiments. In practice, the α-$\mathbf{B}^h$ experiment uses on average 0.04 % more observations than the $\mathbf{B}^s$

Table 8: Mean value of the relative change (in %) in the normalised initial cost function at the start of each assimilation cycle with respect to the initial cost function of the $\mathbf{B}^s$ experiment. In purple, the maximum value across the experiments, and in green the minimum value.

	$\mathbf{B}^s$	α-$\mathbf{B}^h$	$\mathbf{B}^o$
minimum	**100.00**	**98.32**	98.84
maximum	**100.00**	101.30	**102.55**
mean	100.00	**99.50**	**100.81**
median	100.00	**99.50**	**100.77**

Table 9: Same as Tab. 8, but for the mean only and by observation type. For the satellite data, the separation by sensor is also presented.

Observation type	$\mathbf{B}^s$	α-$\mathbf{B}^h$	$\mathbf{B}^o$
Land stations and ships	100.00	**99.69**	**100.50**
Aircraft data	**100.00**	**99.09**	99.57
Atmospheric motion winds	100.00	**99.82**	**100.11**
Drifting Buoys	**100.00**	100.01	**100.38**
Radiosondes	**100.00**	**98.90**	99.82
Balloons and profilers	**100.00**	**99.37**	99.88
Satellite sounding data	**100.00**	99.84	**99.45**
AIRS	**100.00**	**99.98**	99.99
AMSUA	**100.00**	99.76	**99.30**
ATMS	**100.00**	99.24	**95.93**
CRIS	100.00	**99.91**	**100.11**
IASI	**100.00**	99.94	**99.91**
Scatterometer	**100.00**	**99.08**	99.46
Limb observations	**100.00**	**98.97**	99.52
Ground-based precip. composites	100.00	**98.86**	**100.24**
All-sky satellite radiances	**100.00**	**99.26**	99.52
AMSR2	**100.00**	**98.96**	99.38
AMSUA	**100.00**	**98.98**	99.15
GMI	**100.00**	**99.35**	99.61
MHS	**100.00**	**98.96**	99.26
MWHS2	**100.00**	**99.72**	99.80
MWRI	**100.00**	99.32	**99.25**
SAPHIR	100.00	**99.69**	**100.11**
SSMI	**100.00**	**99.01**	99.40
WINDSAT	**100.00**	**99.01**	99.16

experiment (mainly ground-based precipitation composites and satellite sounding data) and the $\mathbf{B}^o$ experiment uses 1.2 % less observations (mainly satellite sounding data) as already presented in Sec. 4.2.1 (Tabs. 3 and 6). To account for that change in the usage of observations, we divide the value of the cost function by the number of used observations. Then, we divide the normalised cost function of the α-$\mathbf{B}^h$ and $\mathbf{B}^o$ experiments by the normalised cost function of the $\mathbf{B}^s$ experiment for each cycle.

The statistics of the relative change of the normalised initial cost function is presented in Tab. 8. The mean values of the change in the initial cost function is smaller for the α-$\mathbf{B}^h$ than the static one by 0.5 %. This indicates a better first-guess fit-to-observation for the α-$\mathbf{B}^h$ experiment on average. The α-$\mathbf{B}^h$ experiment is nonetheless sometimes worse for some assimilation cycles as the maximum change is larger than 100 % by 1.3 %.

The $\mathbf{B}^o$ experiment has a worse first-guess fit-to-observation than the $\mathbf{B}^s$ experiment by 0.8 % on average. If we split this global diagnostic by observation type, then the $\mathbf{B}^o$ experiment is worse mainly for data close to the surface (land stations and ships, drifting buoys), ground-based precipitation composites and AMVs (Tab. 9). For all satellites observations, aircraft data, and radiosondes the $\mathbf{B}^o$ experiment is on average better than the $\mathbf{B}^s$ experiment, with few variations depending on the instrument. The best improvements comes from the ATMS instrument with a reduction of the statistic by about (-4. %). Yet, for almost all observation types but the satellite sounding data (and to a lesser extend the drifting buoys), the α-$\mathbf{B}^h$ experiment is the best. The largest improvements from the α-$\mathbf{B}^h$ experiment are found for the ground-based precipitation composites (-1.14 %) the radiosondes (-1.10 %) and the limb observations (-1.03 %).

The values of the standard deviation of the first-guess departure confirm these results. For the limb observations, the standard deviation of the first-guess departure with respect to the GPS-RO data is lower by about 1 % for the α-$\mathbf{B}^h$ experiment compared to the $\mathbf{B}^s$ experiment for most altitudes (Fig. C.6, page 40). The largest improvement comes from the southern hemisphere above 8 km. This result is reinforced by the comparison with temperature measurements from radiosondes for which the standard deviation of the first-guess departure is also on average lower by about 1 % for the α-$\mathbf{B}^h$ experiment (Fig. C.7, page 40). These data feature an even larger improvement for the polar regions by more than 2 % between 300 hPa and 500 hPa. Concurrently, the $\mathbf{B}^o$ experiment is worse than the static experiment in the Antarctic region and neutral to sightly better elsewhere.

The radiosondes also provide specific humidity measurements. The standard deviation of the first-guess departure with respect to them is on average similar for the α-$\mathbf{B}^h$ experiment and larger by up to 3 % for the $\mathbf{B}^o$ experiment, both compared to the $\mathbf{B}^s$ experiment (Fig. C.8, page 40). The degradation of the $\mathbf{B}^o$ experiment comes from the tropics and from the high latitude regions, but not from polar regions. In particular, in the Antarctic, the $\mathbf{B}^o$ experiment has the lower standard deviation at 300 hPa. For all other pressure levels, the α-$\mathbf{B}^h$ experiment performs better.

These results for specific humidity are opposite to the ones obtained when the first-guesses are compared to the ATMS instrument (Fig. C.10, page 41), for which the channels 18 to 22 are sensitive to tropospheric humidity. For these channels, the $\mathbf{B}^o$ experiment improves the standard deviation by up to 4 % when compared to the $\mathbf{B}^s$ experiment, and for all regions. In the meantime, the α-$\mathbf{B}^h$ experiment improves the standard deviation by just 1 %. A similar signal is also found for the infrared instruments on board geostationary satellites.

The last statistics we examine in more details is the standard deviation of the first-guess departure with respect to the wind speed. Globally, the α-$\mathbf{B}^h$ experiment has the lowest value of standard deviation for all pressure levels, but with a difference less than 1 % compared to the $\mathbf{B}^s$ experiment (Fig. C.5, page 39). The $\mathbf{B}^o$ experiment is marginally better than the $\mathbf{B}^s$ experiment. Again, the largest improvement comes

from the α-$\mathbf{B}^h$ experiment in the polar regions with values of the standard deviation reduced by 1.5 % to 2 % compared to the $\mathbf{B}^s$ experiment. The improvement comes from both components of wind, but it is sightly stronger for the u component of wind.

4.2.4 *Forecast*

To assess the quality of the forecast from each experiment, we have computed the classic IFS scorecard used at ECMWF. If the reader is not familiar with the scorecard, they could find a detailed description in the ECMWF Newsletter No. 160 (Summer 2019) where IFS Cycle 46r1 is compared against IFS Cycle 45r1. The scores are based on the anomaly correlation (or $ccaf$) and root mean square (or $rmsef$) of the forecast error. The forecast error is computed as the difference between the forecast field and here the analysis field from the same experiment ("own analysis") and valid for the same time as the forecast. The forecast error is also computed as the difference between the forecast field interpolated at the radiosondes locations and the radiosondes measurements.

We also computed the forecast error using the analysis field from operation. Since the resolution of the operational analysis is much higher than the one used for our experiments (TCo 1279 versus TCo 399), the operational analysis can be seen as the best available global estimate of the true state of the atmosphere. We found that using the operational analysis was providing the same information on the forecast error as the information provided by the observations. For that reason, we do not to show the results obtained with the operational analysis.

The scores depict if the forecast error is larger or lower for an experiment A compared to the forecast error of an experiment B. We will simplify the discussion hereafter by assuming that a forecast from experiment A is better than a forecast from experiment B, if the $ccaf$ (or $rmsef$) is larger (or smaller) for the forecast error of experiment A compared to the forecast error of experiment B. In that case, it would be a blue triangle in the scorecard (if the result is statistically significant). The opposite would lead to a red triangle facing down (if the result is statistically significant).

Verification against own analysis

Figure 13 presents the scorecard of the α-$\mathbf{B}^h$ experiment versus the $\mathbf{B}^s$ experiment. The forecast verification against own analysis shows that for almost all parameters and all levels, the α-$\mathbf{B}^h$ experiment is better in the first 3 to 5 days. In the tropics, the scores are better for all forecast range. Elsewhere, the α-$\mathbf{B}^h$ experiment is neutral to worse from day 7 to 10, but with a lower statistical significance.

For the short-range forecast, the general improvement from the α-$\mathbf{B}^h$ experiment has two exceptions. The first one is for the temperature at 850 hPa in the southern hemisphere up to day 1. This increase in root mean square error is located mainly on the edges of the polar front and subpolar belt, sightly north of 60°S in a region with a lot of activity and strong gradients. This increase is not present when the forecast is compared against the operational analysis. This indicates that the diagnostic is affected by residual correlations between the forecast error and the analysis error and does not represent the true forecast error. The second exception is for the relative humidity up to day 1. This affects both the southern hemisphere and the northern hemisphere and affects the whole troposphere (not shown). The comparison against the operational analysis does not present any significant degradation there. This, again, could be linked to remaining correlations between the forecast error and the analysis error.

For the $\mathbf{B}^o$ experiment, the comprehensive interpretation of the results from the verification of the forecast against analysis (and against observation too, Fig. 14) is not the focus of this memo. Instead, we just

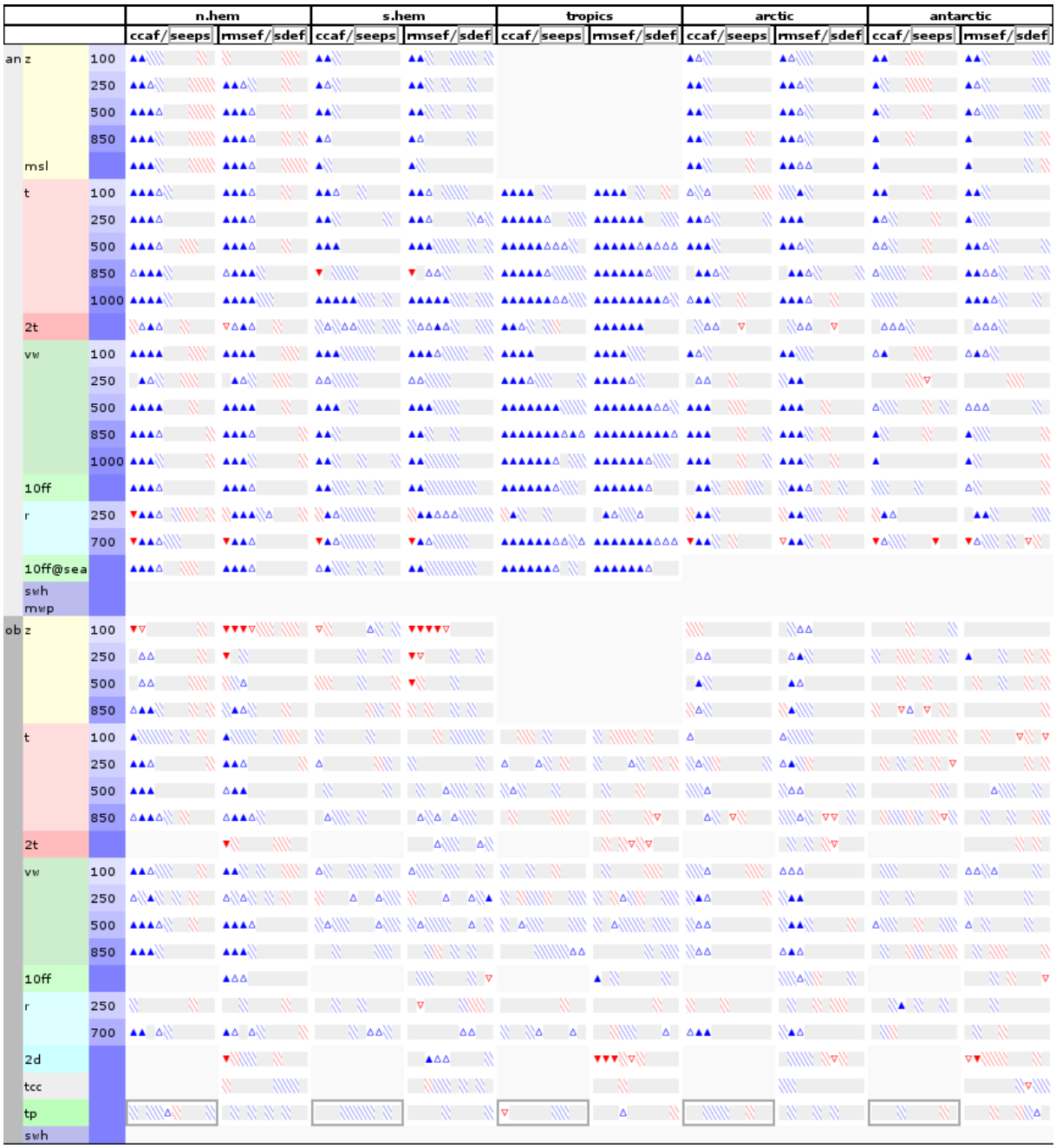

Figure 13: Scorecard of the α-$\mathbf{B}^h$ experiment against the $\mathbf{B}^s$ experiment. There are two columns per region: (left): anomaly correlation or ccaf and (right) root mean square error or rmsef of forecast error. The top part of the table is the comparison against own analysis while the bottom part is the comparison against observation. A blue triangle pointing up means the α-$\mathbf{B}^h$ experiment is better while a red triangle means the α-$\mathbf{B}^h$ experiment is worse. The more intense the color the more significant the statistics. See IFS documentation for more details.

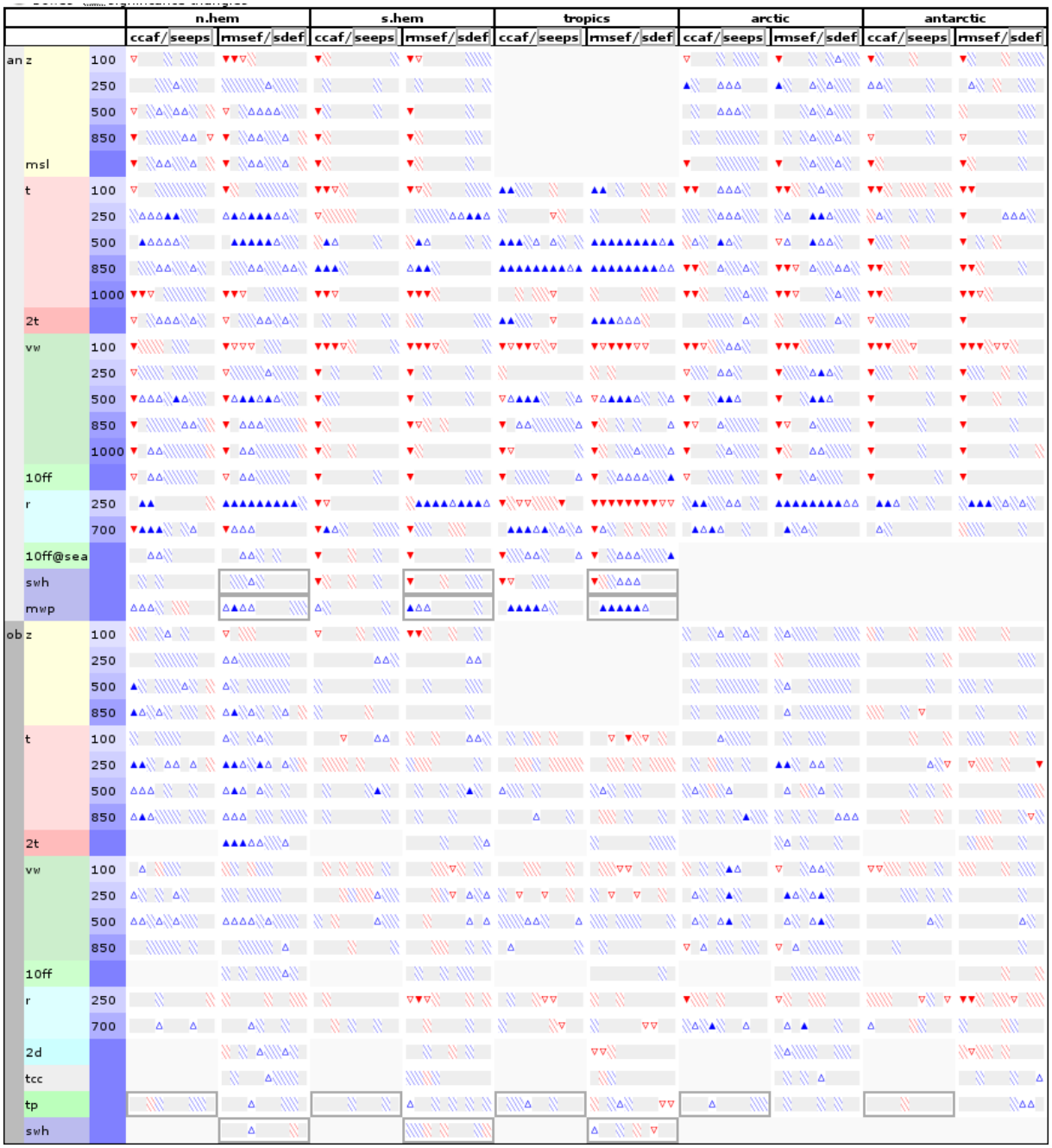

Figure 14: Same as Fig. 13 but for the scorecard of the $\mathbf{B}^o$ *experiment against the* $\mathbf{B}^s$ *experiment.*

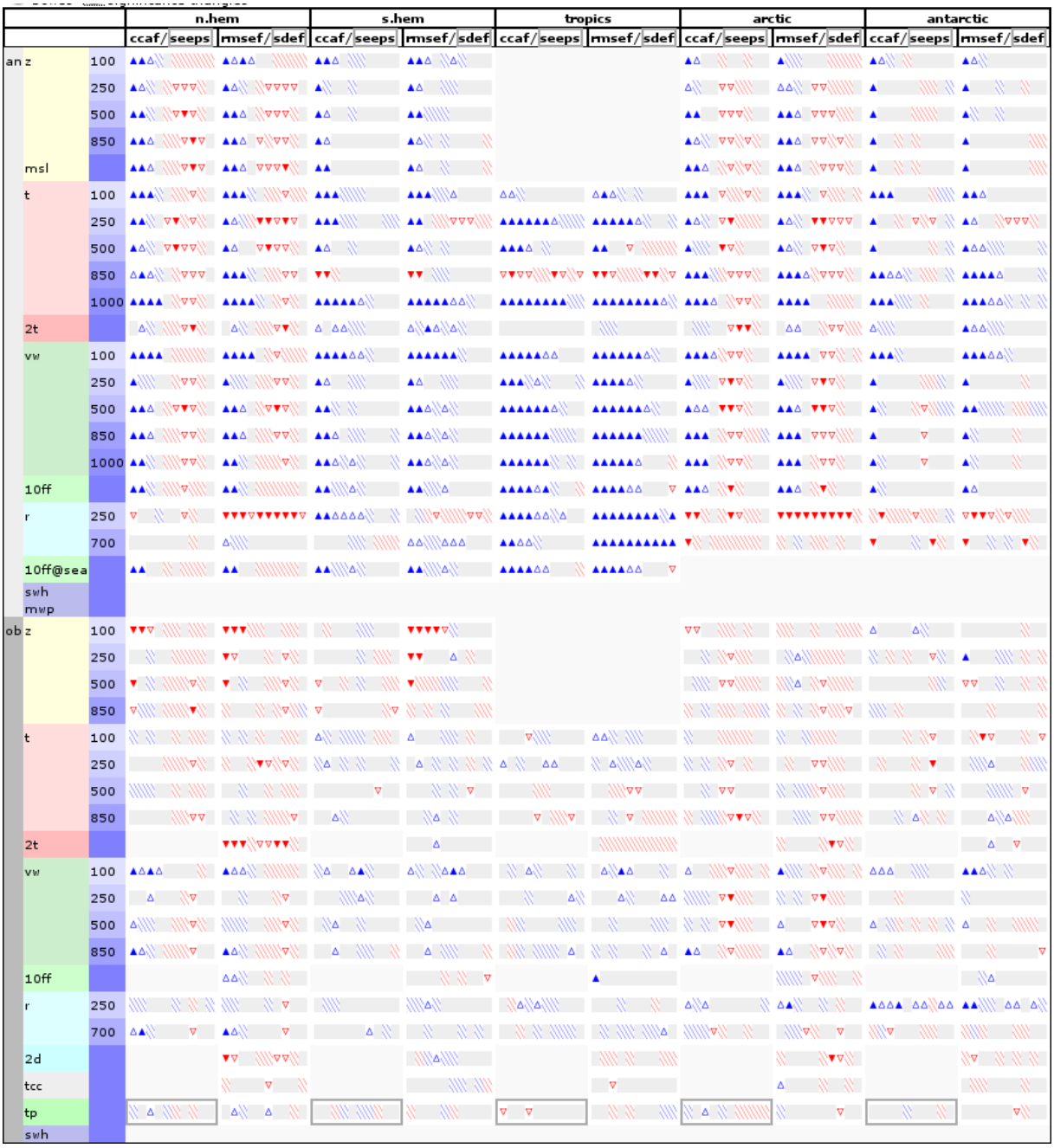

Figure 15: Same as Fig. 13 but for the scorecard of the α-$\mathbf{B}^h$ experiment against the $\mathbf{B}^o$ experiment.

focus on some particular results highlighting the differences with the α-$\mathbf{B}^h$ experiment. For that purpose, we also present the scorecard of the α-$\mathbf{B}^h$ experiment against the $\mathbf{B}^o$ experiment (Fig. 15).

After day 3 to 5, the $\mathbf{B}^o$ experiment produces a better forecast (against own analysis) that the α-$\mathbf{B}^h$ experiment in the northern hemisphere for almost all model variables, and a worse forecast in the southern hemisphere and in the tropics. In the polar regions, the $\mathbf{B}^o$ experiment seems to provide a better forecast in the Arctic but worse in the Antarctic. There are few exceptions to this general behaviour. The first one comes from the relative humidity at 250 hPa with a better forecast from the $\mathbf{B}^o$ experiment for all regions but the tropics. Both experiments nonetheless perform better than the $\mathbf{B}^s$ experiment, except for the $\mathbf{B}^o$ experiment in the tropics. There, the α-$\mathbf{B}^h$ experiment produces the best forecast overall, but for temperature at 850 hPa where the $\mathbf{B}^o$ experiment performs the best.

Verification against observations

When compared to observations, the forecast from the α-$\mathbf{B}^h$ experiment is essentially better than the one from the $\mathbf{B}^s$ experiment only up to day 3 to 5 for all atmospheric variables. For the surface parameters, the α-$\mathbf{B}^h$ experiment is better only in the southern hemisphere and significantly worse for the 2 m dew-point and 2 m temperature in the tropics. For the longer range forecast, there are few significant changes in the forecast skill. The two particular cases are the wind in the southern hemisphere where the forecast from the α-$\mathbf{B}^h$ experiment is better for all levels but for the 10 m wind. The 10 m wind forecast is also significantly worse in the Antarctic.

A significant degradation could be noticed for the geopotential at 100 hPa in the α-$\mathbf{B}^h$ forecast for both hemispheres. This is unexpected since the hybrid formulation of the background error stops just above the tropopause and, above, the formulation is the same as for the $\mathbf{B}^s$ experiment. This could come from some tropospheric-stratospheric exchanges or due to a change in the surface pressure. This will have to be further investigated in the future.

When the forecast error of the α-$\mathbf{B}^h$ experiment against observations is compared to the one of the $\mathbf{B}^o$ experiment, we have a similar picture as when the forecast error is assessed using the own analysis. The forecast for the first few days is generally better for the α-$\mathbf{B}^h$ experiment and, later on, the forecast from the $\mathbf{B}^s$ experiment becomes better. The wind forecast in the southern hemisphere and in the tropics is the exception from this behaviour and is better for the α-$\mathbf{B}^h$ experiment for the whole forecast range and for all levels. The relative humidity forecast at 250 hPa is also better for the α-$\mathbf{B}^h$ experiment for the whole forecast range. The 2 m dew-point and 2 m temperature forecasts from the $\mathbf{B}^o$ experiment are better for all forecast range.

4.2.5 Interaction with VarBC

The change in the analysis between the experiments also depends on the change in the observations usage for each experiment. We already explained that the observations used in the assimilation analysis are first selected using a screening based on the first-guess state. The used observations thus differ from one experiment to the other. Then, each experiment could also make a different usage of certain observations that are bias-corrected. The bias-correction of these observations is indeed analysed within the 4D-Var, together with the initial state, and may differ between the experiments.

Over the period we analysed, there is no significant change in the bias correction provided by the α-$\mathbf{B}^h$ experiment and the one provided by the $\mathbf{B}^s$ experiment. For example, the difference between the bias correction for the AMSU-A instrument is always lower than 0.01 K, with the maximum values found

for the channels 5 to 7, channels sensitive to the tropospheric temperature. On the other hand, the difference of the bias correction between the $\mathbf{B}^o$ experiment and the $\mathbf{B}^s$ experiment for the same the AMSU-A instrument could reach 0.07 K, with the maximum values found for the channels sensitive to the stratosphere.

5 Conclusion

A new hybrid formulation for the background errors was proposed in Massart (2018) and tested in simple configurations. The purpose of the present work was to evaluate this new **B**-formulation against a static **B**-formulation and against the current **B**-formulation in a more realistic configuration. The comparison against the static formulation aims at assessing the impact of flow-dependent background errors. The comparison against current formulation aims at assessing if the new formulation could complement the current one and help improving the forecast scores.

Before performing the evaluation of the new hybrid formulation, this formulation had to be adapted to the current implementation of ECMWF 4D-Var. The analysis is indeed the result of an incremental 4D-Var that successively minimises linear versions of the cost function. The resolution of the linear version of the model used in the analysis increases with each outer loop. We demonstrated that the new hybrid formulation can be used together with an incremental 4D-Var, but one has to consider the change of resolution between outer loops. The change of resolution means that the fields derived from the ensemble of data assimilation (EDA) and used in the new hybrid formulation have to be interpolated at the inner loop resolution for each outer loop. This also means that the localisation length-scale should be adapted to the inner loop resolution.

We assessed the impact of a resolution-dependent localisation length-scale on the analysis. We demonstrated that this influences the analysis result compared to a constant localisation length-scale. At this stage, we are not able to conclude if having resolution-dependent localisation length-scale is beneficial or not. Our formulation of the localisation is currently simple with a constant value in space for the length-scale. We plan to develop a spatially and time-varying localisation formulation in the near future and, then, we will reassess the benefits of a resolution-dependent localisation length-scale.

We compared three experiments to evaluate the new hybrid formulation. The first experiment is based on the current formulation of the background errors and referred to as $\mathbf{B}^o$. The second experiment is based on the new hybrid formulation and uses the forecast fields from the same EDA as the one indirectly used in the $\mathbf{B}^o$ experiment in order to have a fair comparison. This second experiment is referred to as α-$\mathbf{B}^h$. It is using an hybrid weight of 50 % and the static part was built as follows. The variances of the static background errors are the average of the variances of the $\mathbf{B}^o$ experiment over the period of interest. The correlations are the climatological ones used in the $\mathbf{B}^o$ experiment too. We finally designed the last experiment as the one using the static background errors and referred to it as $\mathbf{B}^s$. This means that, on average, the three experiments have the same background error variances.

We compared the three experiments over a winter season, from November 2107 to February 2018 (4 months). The main difference is that the α-$\mathbf{B}^h$ analysis seems less able to correct errors in the large and medium scales probably due to an effective background error covariance that is reduced by the localisation process. The α-$\mathbf{B}^h$ analysis is smoother than the other ones, despite having more energy in the scales above wavenumber around 100. This is mainly the case for divergence (and to a lesser extend relative humidity) where the α-$\mathbf{B}^h$ analysis has less energy in the large to medium scales, up to the wavenumber around 100. In the meantime, the $\mathbf{B}^o$ analysis has more energy in these scales and is more active. We believe that lower effective covariance and the resulting smoothness of the α-$\mathbf{B}^h$ experiment

explains why it converged faster than the two other experiments and why it has a lower condition number.

The first-guess derived from the α-$\mathbf{B}^h$ analysis is on average closer to the observations than the one derived from the two other experiments, by 0.5 % compared to $\mathbf{B}^s$ and by 1.3 % compared to $\mathbf{B}^o$. This is true for all observation types except for the satellite sounding data for which the $\mathbf{B}^o$ experiment performs better. For the satellite, the main difference is for the ATMS instrument for which the $\mathbf{B}^o$ first-guess is closer to the observations by an average of 3.3 % compared to α-$\mathbf{B}^h$ first-guess. This could be one of the reasons behind the difference behaviour between the experiment for the medium-range forecast.

When compared to their own analysis, the forecast from the α-$\mathbf{B}^h$ experiment is indeed better than the forecast from the other experiments, but only up to day 3 to 5. Then the forecast from the $\mathbf{B}^o$ experiment tends to be the best. There are few exceptions though also highlighted by the comparison with observations. For all forecast range, the α-$\mathbf{B}^h$ experiment does produce the worse forecast for the surface parameters (2 m dew-point and 2 m temperature and to a lesser extent 10 m wind), but for the southern hemisphere. The wind forecast is better for all forecast range in the southern hemisphere and tropics when it is derived from the α-$\mathbf{B}^h$ experiment. The relative humidity forecast from α-$\mathbf{B}^h$ experiment is also in general the best.

The effective background error covariance and therefore the smoothness of the α-$\mathbf{B}^h$ analysis are very much dependent of the localisation length-scale. A smaller localisation length-scale may decrease the effective covariance and prevent the analysis to fit the spatial variability present in the observations and thus may provide a smoother analysis. A larger location length-scale may increase the effective covariance and allow the analysis to have more spatial variability as the analysis would fit better the observations. By influencing the distribution of the energy of the increments in the spectral space, the localisation length-scale consequently influences the forecast. For all these reasons, we will pay more attention in the future to our formulation for the localisation. In particular, we will investigate the benefits of having a space-dependent localisation length-scale and then a time-dependent one by updating it at every assimilation cycle. This work will be carried on together with moving toward using perturbations derived from the currently operational 50 members EDA.

The smoothness of the α-$\mathbf{B}^h$ analysis can also be a consequence of the implementation choices of the α-$\mathbf{B}^h$ formulation. Indeed, the increments of the model variables derived from the ensemble-based part of the background error covariances are intimately linked together through the α variables. This reduces the degree of freedom for each variable. We should assess in the future the cost and benefits of having different α variables for each model variable. Moreover, the α variables are supposed to be valid for all model levels in the troposphere. This also reduces the degree of freedom of the analysis. On the other hand, this seems not to be an issue as there is no clear difference in behaviour on the vertical for the α-$\mathbf{B}^h$ analysis when compared to the $\mathbf{B}^s$ analysis.

Finally, the forecast of the surface parameters are mostly worse for the α-$\mathbf{B}^h$ experiment. The diagnostics for the localisation length-scale suggest that the length-scale may differ significantly between the lowest model levels and the mid-troposphere. It may be necessary, in the future, to restrict the new hybrid formulation to the free troposphere and, eventually, to have a decoupled hybrid formulation for the planetary boundary layer.

6 Acknowledgements

The author would like to thank Elias Holm and Massimo Bonavita for their careful reading of draft versions of this document and for their helpful comments and recommendations.

The author would also like to thank Mark Rodwell for helping interpreting and summarising the outputs of Diagnostic Explorer.

The discussion of the results also benefited from the expertise of various persons from the Earth System Assimilation Section at ECMWF.

References

Berre, L., Arbogast, E., Ménétrier, B., and Desroziers, G. (2017). Change of variable applied to mass and wind fields for covariance localisation. WMO CAS/JSC WGNE Blue Book, Edited by J. Côté.

Bonavita, M., Isaksen, L., and Hólm, E. (2012). On the use of EDA background error variances in the ECMWF 4D-Var. *Q.J.R. Meteorol. Soc.*, 138(667):1540–1559.

Courtier, P., Thépaut, J.-N., and Hollingsworth, A. (1994). A strategy for operational implementation of 4D-Var, using an incremental approach. *Q.J.R. Meteorol. Soc.*, 120:1367–1388.

Desroziers, G., Camino, J.-T., and Berre, L. (2014). 4DEnVar: link with 4D state formulation of variational assimilation and different possible implementations. *Q.J.R. Meteorol. Soc.*, 140(684):2097–2110.

Fisher, M. and Andersson, E. (2001). Developments in 4D-Var and Kalman Filtering. ECMWF Technical Memorandum No. 347.

Gaspari, G. and Cohn, S. (1999). Construction of correlation functions in two and three dimensions. *Q.J.R. Meteorol. Soc.*, 125(554):723–757.

Massart, S. (2018). A new hybrid formulation for the background error covariance in the IFS: implementation aspects. ECMWF Technical Memorandum No. 832.

Ménétrier, B. and Auligné, T. (2015). Optimized localization and hybridization to filter ensemble-based covariances. *Monthly Weather Review*, 143(10):3931–3947.

Talagrand, O. (1997). Assimilation of Observations, an Introduction. *Journal of the Meteorological Society of Japan*, 75(1B):191–209.

Appendix A Building a climatology of the background errors

In the IFS, the background error covariance matrix is decomposed into three main operators that can be applied sequentially: the background error correlations of the unbalanced variables ($\mathbf{C}$), the background error standard deviation (Σ^b) and the balance operator $\mathbf{K}$,

$$\mathbf{B} = \mathbf{K}\mathbf{S}_J\Sigma^b\mathbf{C}\Sigma^{b^T}\mathbf{S}_J^T\mathbf{K}^T . \tag{17}$$

The operation $\mathbf{S}_J$ performs the spectral transform at the highest wavenumber N_J for the variables of the unbalanced control vector that are in spectral space.

In the new hybrid formulation for the background errors, it is expected to combined the ensemble-based covariance matrix $\mathbf{B}^e$ with a static covariance matrix $\mathbf{B}^s$ (see Massart (2018) for more details). In order to have consistent background errors for all the experiments presented in this document (see Sec. 4), we decided to build the static covariance matrix $\mathbf{B}^s$ as the average of the flow dependent ensemble-based background errors over the experimental period (1 November 2017 to 1 March 2018). Let us assume that we have K assimilation cycles of 12 h during this period and that for each cycle k we have computed the flow-dependent standard deviation of the background error Σ_k^b. Then we computed the static background error standard deviation Σ^s as the square-root of the mean variance over the K assimilation cycles,

$$\Sigma^s = \sqrt{\frac{1}{K}\sum_{k=1}^{K}\Sigma_k^{b^2}} . \tag{18}$$

For the background error correlation matrix, we just used the climatological correlation matrix $\mathbf{C}^s$ used to compute the flow-dependent correlation matrix for each cycle. We did not average the statistical balance operator, so the static background error covariance matrix take the form

$$\mathbf{B}^s = \mathbf{K}\mathbf{S}_J\Sigma^s\mathbf{C}^s\Sigma^{s^T}\mathbf{S}_J^T\mathbf{K}^T . \tag{19}$$

There was an additional step to consider. The background error standard deviation Σ_k^b is computed at a given resolution and a scaling factor α is apply to compensate for the difference in resolution between Σ_k^b and the resolutions of the inner loops. The scaling factor α varies with time and is different for all atmospheric model variables and all levels. It is based in the mean standard deviation per level and per variable. We then computed the mean standard deviation profile per variable over the K assimilation cycles following the same formula as used in Eq. (18).

(a) TL95 *(b) TL159* *(c) TL255*

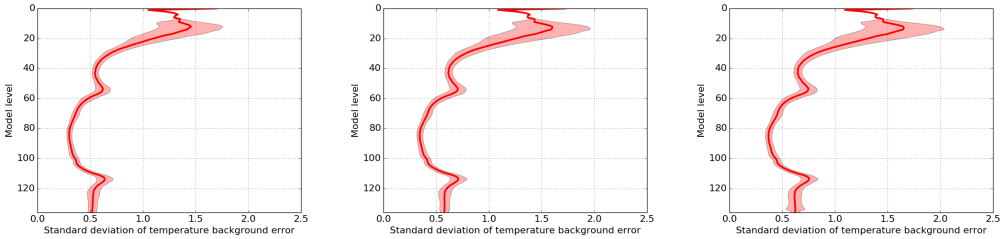

Figure A.1: Example of mean profile of background error standard deviation for temperature (in K*) and for 3 resolutions: (a) TL95, (b) TL159, (c) TL255. The values are obtained from an average over the period 1 November 2017 to 1 March 2018 and the shade area represents the variability (one standard deviation) during this period.*

Appendix B Change in analysis mean fields

B.1 Zonal mean

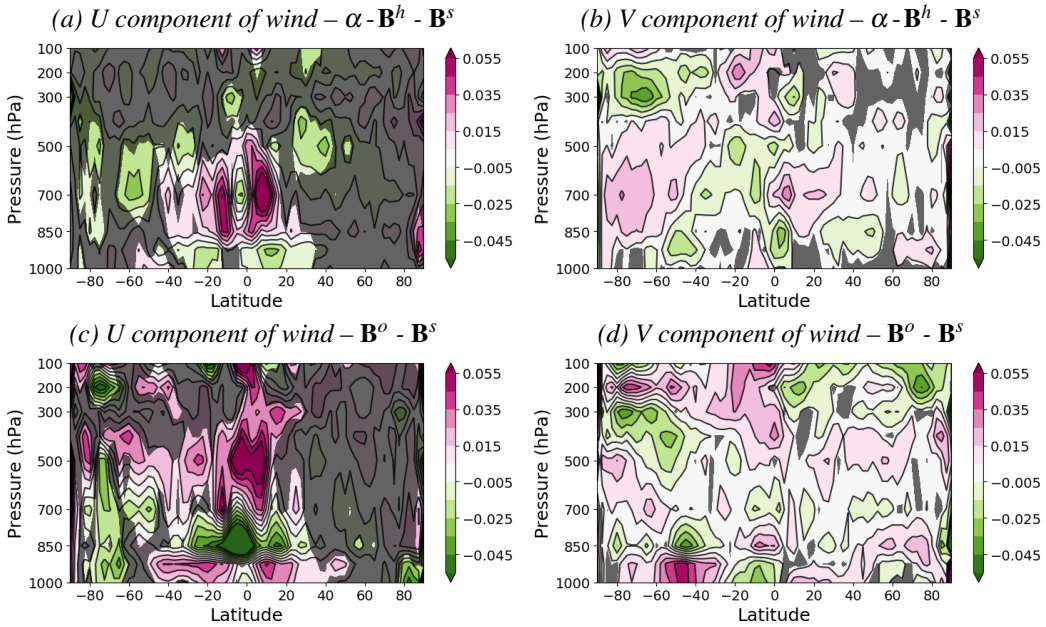

Figure B.1: Change in mean field in u (left) and v (right) component of wind (in $\mathrm{m.s}^{-1}$*) at initial time between the* α*-*$\mathbf{B}^h$ *and the* $\mathbf{B}^s$ *experiments (top) and between the* $\mathbf{B}^o$ *and the* $\mathbf{B}^s$ *experiments (bottom). The grey areas indicate a non statistically significant change.*

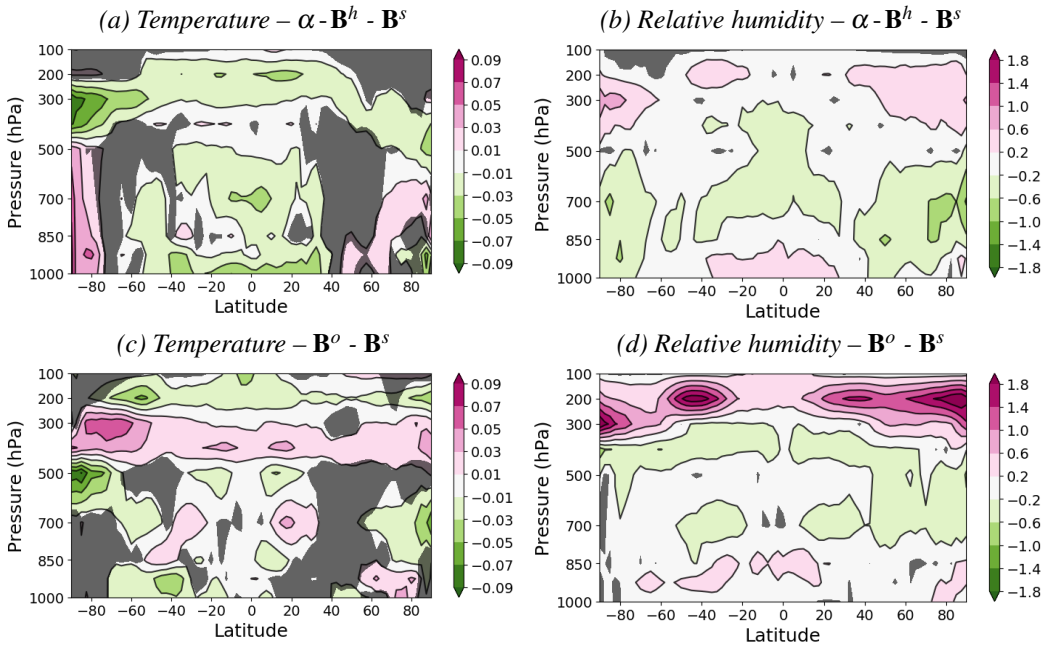

Figure B.2: Same as Fig. B.1 but for temperature (left, in k*) and relative humidity (right, in %).*

B.2 Mean spectra

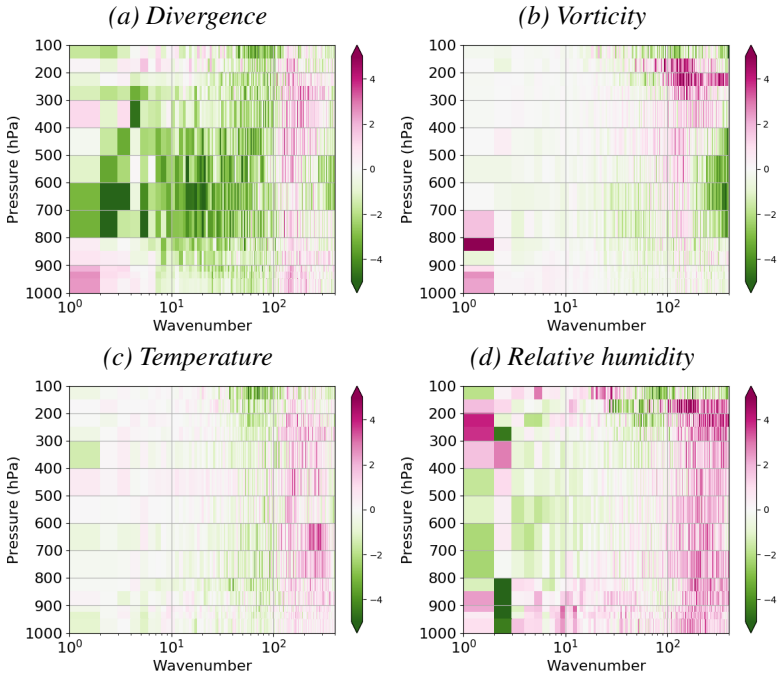

Figure B.3: Spectra of the normalised mean difference (in %) between the analysis of the α-$\mathbf{B}^h$ experiment and the one from the $\mathbf{B}^s$ experiment as a function of pressure. The mean is computed over the month of November 2017. The difference is normalised by the mean spectra of the $\mathbf{B}^s$ experiment.

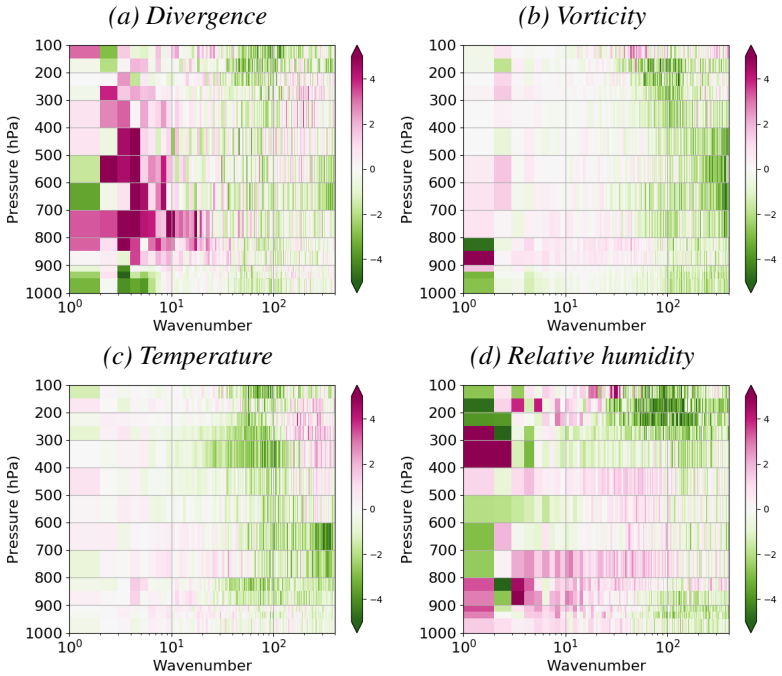

Figure B.4: Same as Fig. B.3, but for the difference between the $\mathbf{B}^o$ and the $\mathbf{B}^s$ experiments.

Appendix C Statistics of the experiments against observations

C.1 Analysis

C.1.1 Wind related observations

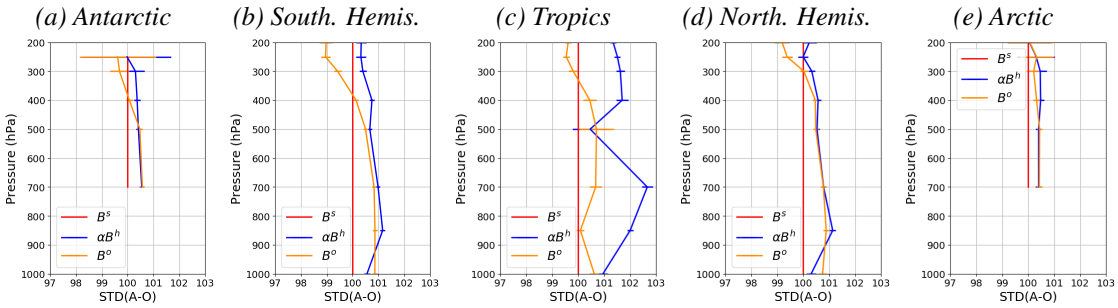

Figure C.1: Normalised standard deviation (in %) of the AMV analysis departure as a function of pressure level and by region, and for all three experiments: (red) $\mathbf{B}^s$, (blue) α-$\mathbf{B}^h$, (orange) $\mathbf{B}^o$. The normalisation is with respect to the $\mathbf{B}^s$ experiment. The error-bars represent the variability of each value.

C.1.2 Temperature related observations

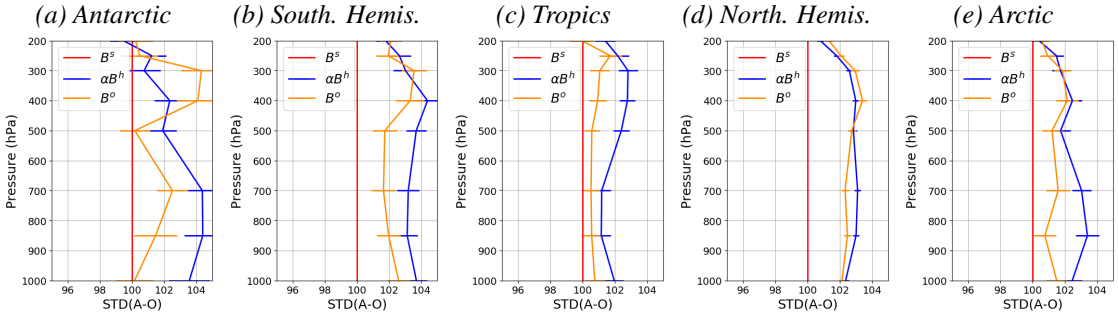

Figure C.2: Same as Fig. C.1 but for the standard deviation of the analysis departure with respect to temperature measurements from radiosondes.

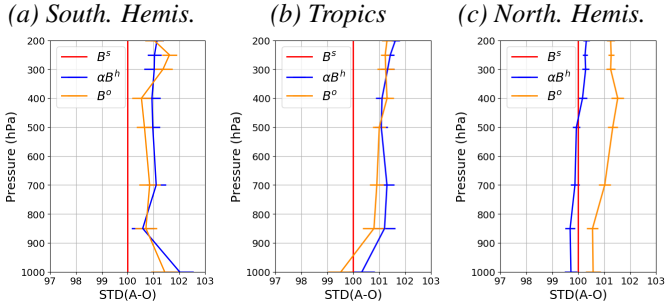

Figure C.3: Same as Fig. C.1 but for the standard deviation of the analysis departure with respect to temperature measurements from aircraft data.

C.1.3 Humidity related observations

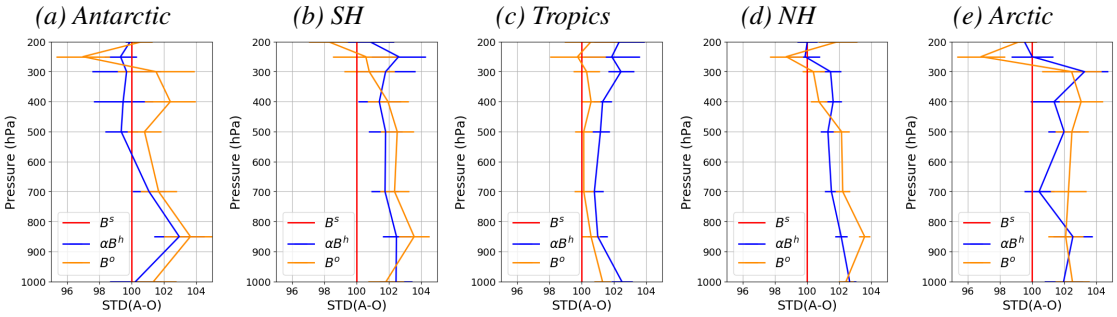

Figure C.4: Same as Fig. C.1 but for the standard deviation of the analysis departure with respect to specific humidity measurements from aircraft data.

C.2 First-guess

C.2.1 Wind related observations

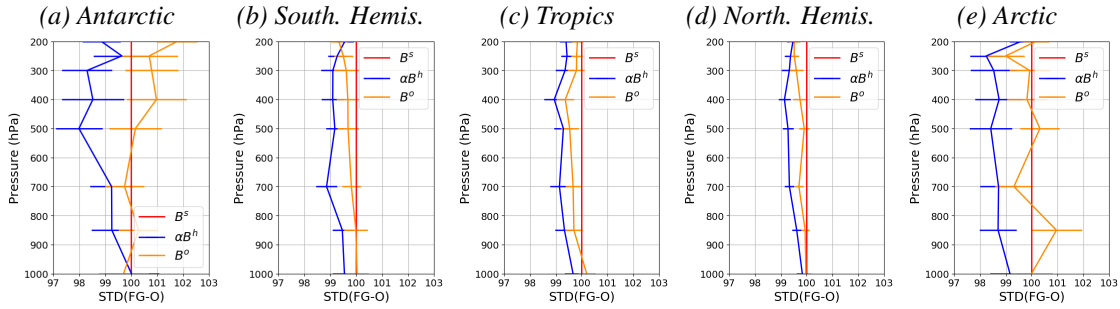

Figure C.5: Normalised standard deviation (in %) of the wind speed first-guess departure compared to radisondes, as a function of pressure level and by region, and for all three experiments. The normalisation is with respect to the $\mathbf{B}^s$ experiment. The error-bars represent the variability of each value.

C.2.2 Temperature related observations

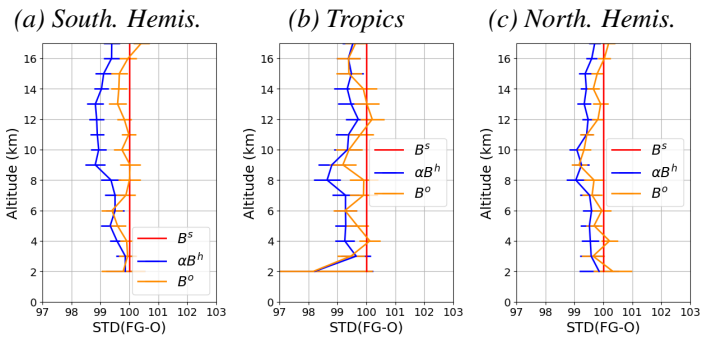

Figure C.6: Same as Fig. C.5 but for the standard deviation of the first-guess departure with respect to specific GPS-RO measurements.

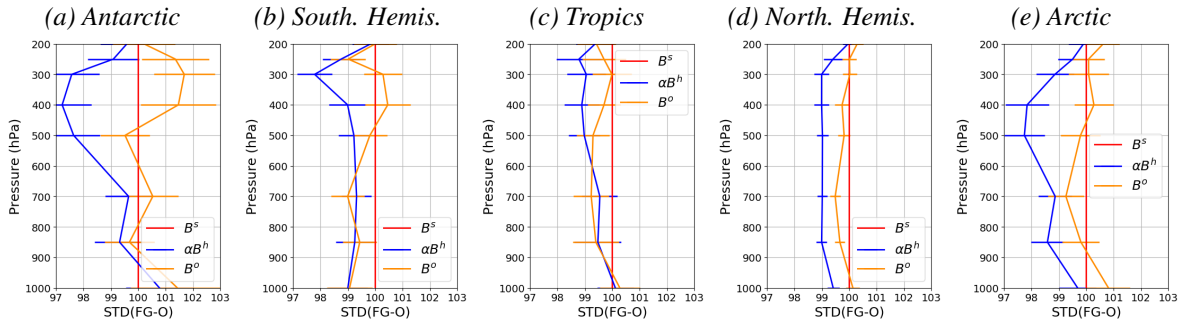

Figure C.7: Same as Fig. C.5 but for the standard deviation of the first-guess departure with respect to temperature measurements from radiosondes.

C.2.3 Humidity related observations

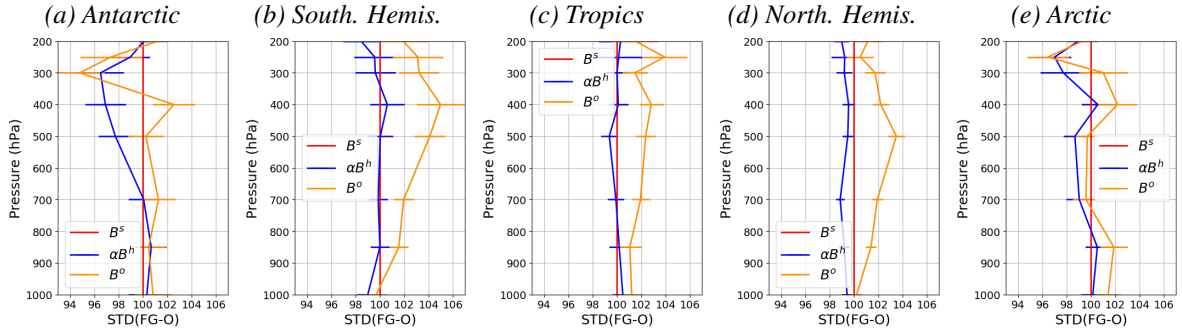

Figure C.8: Same as Fig. C.5 but for the standard deviation of the first-guess departure with respect to specific humidity measurements from aircraft data.

C.2.4 Satellite related observations

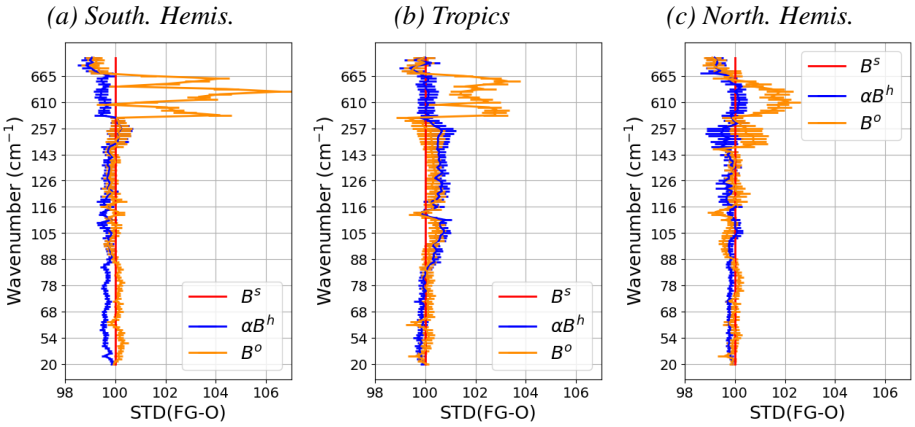

Figure C.9: Same as Fig. C.5 but for the standard deviation of the first-guess departure with respect to CRIS measurements.

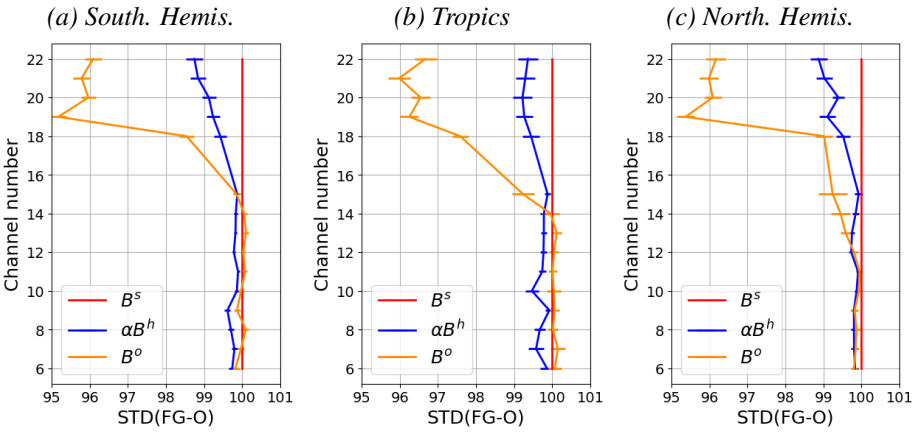

Figure C.10: Same as Fig. C.5 but for the standard deviation of the first-guess departure with respect to ATMS measurements.

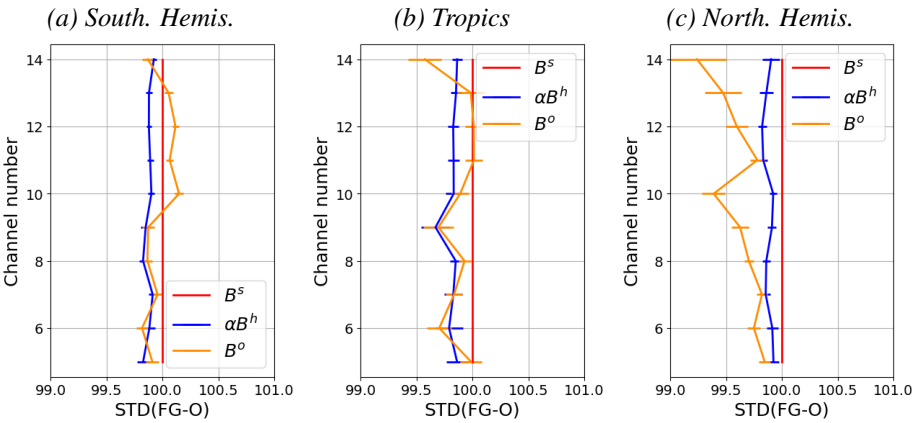

Figure C.11: Same as Fig. C.5 but for the standard deviation of the first-guess departure with respect to AMSUA measurements.