Background error correlation length-scale estimates and their sampling statistics

Olivier Pannekoucke, Loïk Berre, Gérald Desroziers

QUARTERLY JOURNAL OF THE ROYAL METEOROLOGICAL SOCIETY
Q. J. R. Meteorol. Soc. **134**: 497–511 (2008)
Published online in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/qj.212

Background error correlation length-scale estimates and their sampling statistics

O. Pannekoucke*, L. Berre and G. Desroziers
GAME/CNRM (Météo-France, CNRS), Toulouse, France

Abstract: This article presents different formulae to estimate correlation length-scales, and an evaluation of their qualities for practical diagnostic applications. In particular, two new and simple formulae are introduced, which only require the computation of correlation with a single point for a given direction. It is then shown in a 1D heterogeneous context that all formulations lead to similar realistic length-scale values, and that they represent geographical variations rather well.
The estimation of length-scales within a finite ensemble is also studied. While a positive bias occurs when the ensemble size is too small, the standard deviation of the length-scale estimation is shown to be the main influence on the estimation error. The spatial structure of sampling noise is then diagnosed, and effects of spatial filtering techniques on the bias and standard deviation are illustrated.
Finally, an ensemble of perturbed forecasts from a global NWP model is used, showing a real application example.

WARNING : This is a preprint of an article accepted for publication in QUARTERLY JOURNAL OF THE ROYAL METEOROLOGICAL SOCIETY ref: *Q. J. R. Meteorol. Soc.* **134**: 497–508 (2008) see the website for final version http://www.interscience.wiley.com/

KEY WORDS Data assimilation; Diagnosis; Length-scale approximation; ensemble; sampling noise.

Received 28 June 2007; Revised 19 December 2007; Accepted 19 December 2007

1 Introduction

In order to objectively determine initial conditions for numerical weather prediction, modern data assimilation schemes rely on specified error statistics to obtain an approximately optimal combination of observations and a background given by a short-range forecast. This near-optimal analysis is derived from statistical estimation theory. In this framework, the two sets of information are associated with covariance matrices corresponding to their respective errors. The error covariance matrices determine the respective weights given to each piece of information in the analysis. However, the correct specification of those statistics remains a major challenge in data assimilation systems.

The structure of correlation functions is particularly important, as it determines how the observed information is filtered and propagated spatially. Typically, when the background error structure is large scale, the correlation functions are relatively wide. This implies that small scale observed details tend to be filtered out in the analysis step, and that locally observed information is propagated on large spatial distances (Daley, 1991). Diagnostics of the length-scale of background error correlation functions are often used as an approximate indicator of the degree of spatial smoothing. Following the classical definition of a differential length-scale by Daley (1991, p110), the length-scale diagnosis describes the curvature of the correlation functions near their origin. Thus, the smaller the length-scale is, the faster the correlation decreases with distance.

As illustrated by several authors (Hollingsworth, 1987; Bouttier, 1993; Rabier *et al.*, 1998; Ingleby 2001, Belo Pereira and Berre, 2006; Deckmyn and Berre, 2005), this length-scale diagnosis also gives information about atmospheric dynamic (Ingleby, 2001) and data density effects (Bouttier, 1993) on the background error spatial structures. Therefore, it is attractive to be able to diagnose and to interpret length-scales at different locations. This is all the more important as ongoing research is devoted to the representation of existing heterogeneities and anisotropies (e.g. Fisher, 2003; Buehner, 2005).

In this paper, the local length-scales have been approximated by different formulae. A first purpose of the paper is thus to evaluate the ability of these various formulae to diagnose the geographical variations of the local length-scales.

In addition, with the availability of forecast ensembles, it is possible to calculate flow-dependent background error covariances "of the day" (Kalnay, 2002). However, the finite size of the ensemble induces a sampling noise, which is detrimental for the covariance estimation.

*Correspondence to: Météo-France CNRM/GMAP, 42 av. G. Coriolis, 31057 Toulouse Cedex France. e-mail: olivier.pannekoucke@meteo.fr

Regarding correlations, the sampling noise has been studied mostly with respect to long distance correlation values, which were identified as particularly noisy (Houtekamer and Mitchell, 2001). By contrast, relatively little is known about the level of noise in the estimated local length-scales. A second purpose of the current paper is thus to study this sampling noise in length-scale estimates.

The current paper deals with the length-scale as a diagnostic of exisiting correlation estimates. The focus is not on modelling correlation functions on the basis of estimated length-scales, althaugh this is another potential application in the future.

The structure of the paper is as follows. In section 2, different formulae for length-scale are derived from Daley's definition. Experimental results are illustrated in section 3, in a simple 1D analytical framework. Section 4 shows the sensitivity of length-scale estimation to the ensemble size and the spatial structure of sampling noise. Section 5 presents the comparison of two length-scale formulae, in the spherical case, using an ensemble of perturbed forecasts from the French NWP model Arpège.

2 Length-scale formulae

One of the main issues for a data assimilation system is to better specify the background error covariance matrix $\boldsymbol{B} = \mathbb{E}(\boldsymbol{\varepsilon}_b \boldsymbol{\varepsilon}_b^T)$, where $\boldsymbol{\varepsilon}_b$ is the forecast error assumed unbiased. In order to characterize the curvature of the correlation functions near their origin, length-scale diagnosis is often introduced.

The differential length-scale is defined in data assimilation following Daley (1991, p110). The definition is similar to the turbulent microscale. In this section, the Daley length-scale is reviewed, and formulae are derived to approximate it.

The decomposition of covariances into standard deviations and correlations is common e.g. in variational schemes. This is appropriate if standard deviations and correlations do not vary much on scales smaller than the correlation length-scale.

2.1 Daley formula

For a smooth and isotropic correlation function ρ at the origin, the Daley length-scale is given by

$$L_D = \sqrt{-\frac{1}{\nabla^2 \rho(0)}}, \tag{1}$$

in one dimension and $L_D = \sqrt{-\frac{2}{\nabla^2 \rho(0)}}$ in two dimensions. This length-scale is proportional to the turbulent (or Taylor) microscale which is similarly defined. This formula is obtained from a Taylor expansion of the correlation at the origin $\rho(0)$:

$$\rho(\delta x) \approx \rho(0) + \frac{\delta x^2}{2}\frac{d^2\rho}{dx^2}(0) = 1 - \frac{\delta x^2}{2L_D^2}. \tag{2}$$

The isotropic assumption is required in order to ensure the continuity of the second order derivative at 0, *i.e.* $\frac{d^2\rho}{dx^2}(0^-) = \frac{d^2\rho}{dx^2}(0^+)$. A geometrical interpretation of this definition of length-scale is given as the scale for which the tangential parabola at the origin is equal to 0.5. This is illustrated in the top panel of Fig. 1, where a correlation function (solid line) and its tangential parabola at the origin (dashed line) are represented. The length-scale deduced from the above geometrical interpretation is $L_D = 250\,km$, for this particular correlation function. The length-scale is also related to the curvature of the correlation function, at the origin. The radius of curvature of the correlation function at the distance r is defined by $R(r) = \frac{\left(1+\left(\frac{d\rho}{dx}(r)\right)^2\right)^{3/2}}{\frac{d^2\rho}{dx^2}(r)}$. At the origin, $\frac{d\rho}{dx}(0) = 0$ leading to $R(0) = \frac{1}{\frac{d^2\rho}{dx^2}(0)} = -L_D^2$.

Note that the Daley length-scale does not give information about the correlation anisotropy. Moreover, it requires the knowledge of the second order derivative of the correlation function. The calculation of this second order derivative can be rather costly, as ideally it should involve the calculation of the whole correlation function. The next subsections will thus describe convenient approximations of this formula.

2.2 Belo Pereira-Berre formula

Belo Pereira and Berre (2006) (hereafter noted B&B) have proposed a relatively costless formula for the computation of length-scale. Under local differentiability and local homogeneity assumptions, the variance of the spatial derivative of the forecast error can be approximated by $\left(\sigma(\partial_x \varepsilon_b(x))\right)^2 = \left(\partial_x \sigma(\varepsilon_b(x))\right)^2 - \left(\sigma(\varepsilon_b(x))\right)^2 \partial_x^2 \rho(0)$, where $\partial_x = \frac{\partial}{\partial x}$ is the derivative along the coordinate. From the Daley length-scale definition, it follows:

$$L_{B\&B} = \sqrt{\frac{\left(\sigma(\varepsilon_b(x))\right)^2}{\left(\sigma(\partial_x \varepsilon_b(x))\right)^2 - \left(\partial_x \sigma(\varepsilon_b(x))\right)^2}}, \tag{3}$$

where $\sigma\left(\varepsilon_b(x)\right)$ is the standard deviation of $\varepsilon_b(x)$. This formula method requires the computation of forecast error standard deviation, its gradient and also the standard deviation of the gradient of forecast error. In the case of a periodic domain, the computation of the gradient can be done either in grid-point space or in spectral space.

2.3 Parabola-based and Gaussian-based formula

As suggested by equation (2), a direct discretization of the Laplacian appearing in Eq. (1) leads to a simple expression of the length-scale

$$L_{Pb} = \frac{\delta x}{\sqrt{2\left(1-\rho(\delta x)\right)}}. \tag{4}$$

This length-scale is called hereafter the parabola-based length-scale (Pb). It is based on the approximation of the correlation function by a parabolic function, as represented in Fig. 1. As suggested by the example shown in

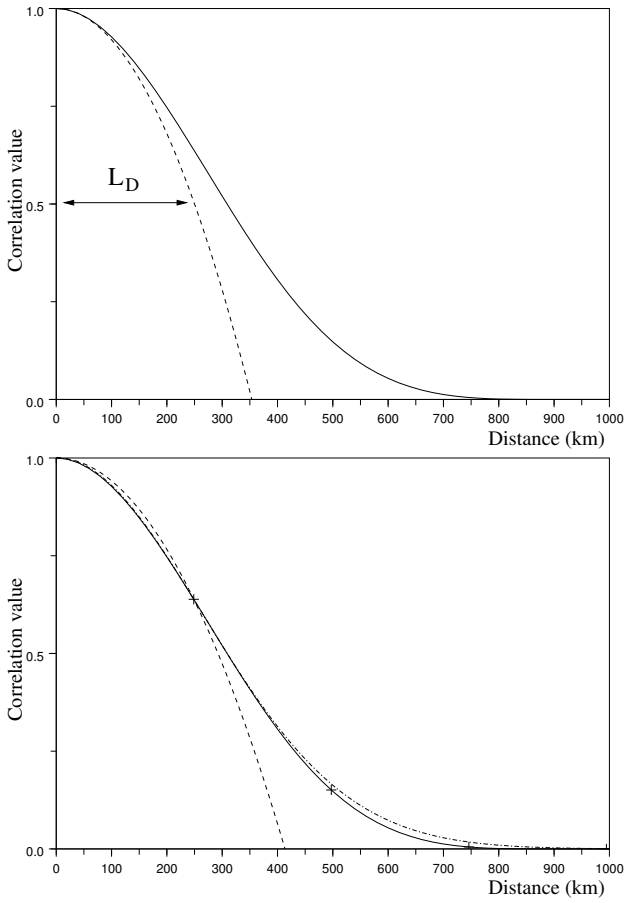


Figure 1. Top panel : Gaspari and Cohn (see section 3.1) correlation function (solid line) and its tangential parabola (dashed line). Bottom panel : Parabolic (dashed line) and Gaussian (dash-dotted line) approximations at the origin of the correlation function (solid line), determined by its value for $\delta x = 248km$, on a regular grid (crosses).

the bottom panel of Fig. 1, for some separation distances (those smaller thant the chosen distance δx), the parabolic function may decrease less quickly (from the origin to the chosen distance δx) than the true correlation function.

This suggests that the quality of the parabolic length-scale approximation may depend on the quality of the correlation function approximation and on the considered separation distance δx. Experiments indicate that the sensitivity to the choice of δx is relatively small, and that using a small value for δx provides a somewhat more accurate estimate of the length-scale. In this paper, δx corresponds to the resolution of the grid (*i.e.* the smallest possible δx).

In order to study this sensitivity to the correlation shape approximation, it is thus interesting to consider another analytical model of the correlation function ρ near the origin. By approximating the correlation at the origin by a Gaussian, the following equation is obtained: $\rho(\delta x) = exp(-\frac{\delta x^2}{2\,L_D^2})$. Inverting this equation to extract the length-scale formulation, associated to correlation at distance δx, brings

$$L_{Gb} = \frac{\delta x}{\sqrt{-2\ \ln\ \rho(\delta x)}}. \tag{5}$$

This length-scale is called hereafter the Gaussian-based length-scale (Gb). This approximation of the length-scale computation is easy to implement in real applications and costless. The bottom panel of Fig. 1 illustrates the Gaussian approximation at the origin of the discretized correlation function.

Note that when the correlation is close to one, then both Parabola-based and Gaussian-based length-scales are equal. Let $\eta = 1 - \rho$, then a Taylor expansion leads to $L_{Pb} = L_{Gb} = \frac{\delta x}{\sqrt{2\eta}}$.

2.4 Directional length-scale

Formulae (4) and (5) can be defined along an arbitrary direction as follows. Let $\boldsymbol{\delta x}$ be the displacement in a direction $\boldsymbol{u} = \frac{\boldsymbol{\delta x}}{|\boldsymbol{\delta x}|}$ of the domain (circle, plane, 2D-sphere, 3D-sphere,...). Then the vectorial parabola-based and Gaussian-based length-scale are thus defined by replacing δx by $\boldsymbol{\delta x}$ in equation (4) and (5). Thus it offers a characterization of the correlation for different directions.

Similarly, formula (1) and (3) can be defined directionally for an anisotropic correlation function. For equation (1), it consists in replacing $\Delta\rho(0)$ by $\frac{\partial^2\rho}{\partial \boldsymbol{u}^2}(0^+) = \lim_{t\to 0^+} 2\left\{\rho(t\,\boldsymbol{u}) - 1\right\} t^{-2}$, which is the second order derivative, calculated in the oriented direction $\boldsymbol{u}$, of the anisotropic correlation function. For equation (3), the directional length-scale is obtained by calculating the gradients $\partial_{\boldsymbol{u}}\varepsilon_b$ and $\partial_{\boldsymbol{u}}\sigma(\varepsilon_b)$, where $\partial_{\boldsymbol{u}}$ is the derivation along $\boldsymbol{u}$.

It should be noted that these length-scales can be calculated whether the domain is bounded or not. Thus such formulations are suitable in oceanography or for a limited area model, as well as for a global meteorological model.

In the particular case of a 1D domain, one can define a directional parabola-based left length-scale as $L_{Pb}(-\delta x)$ and a right length-scale as $L_{Pb}(+\delta x)$. A similar definition is given for the directional Gaussian-based length-scale. Thereafter, the left directional length-scale is designed by a superscript $^-$ and the right one by the superscript $^+$. Note that the ratio $\frac{L^+}{L^-}$ is an indicator of anisotropy.

2.5 Other length-scale formulae

The length-scale can be approximated in other ways, by considering various analytical expressions for correlation

$$\rho(\delta x) = f(\delta x, L_D). \tag{6}$$

The main constraint is that the formula for f has to be invertible. Thus the length-scale can be deduced from the correlation value, associated to a particular grid, with $L_D = f^{-1}\left(\delta x, \rho(\delta x)\right)$. The choice of a particular relation

DOI: 10.1002/qj

between correlation and length-scale may arise from estimated correlation functions. It might depend on the physical field, or on the model used to represent the correlation function in the system.

For instance, if a SOAR function (Daley, 1991, p117) is a good model to approximate the correlation function, then one has to invert Eq. (6), with $f(\delta x, L) = (1 + \frac{\delta x}{L})e^{-\frac{\delta x}{L}}$. This inversion can be achieved by using a Newton algorithm to resolve $F(L) = 0$, with $F(L) = \rho(\delta x) - f(\delta x, L)$ where δx and $\rho(\delta x)$ are given.

Moreover, it can be noticed that such development may be applied on more complex diagnosis in 1D, 2D and 3D.

In the following, the 1D circle and the 2D sphere will be considered in order to illustrate the theory.

3 Application in a 1D analytical heterogeneous framework

3.1 A simple 1D analytical framework

Following Pannekoucke *et al.* (2007), a simple 1D analytical framework is considered to evaluate the quality of the various formulations of length-scale explored in this paper. In this framework, the geographical domain is supposed to be the equatorial circle of radius a, and the coordinate $\frac{x}{a}$ is the angle of the geographical position, varying from $0\,^\circ$ to $360\,^\circ$. On this circle, only one field is considered. A homogeneous Gaussian correlation tensor, is produced following $B_h(x,y) = e^{-\frac{(x-y)^2}{2L_H^2}}$, where x and y are two points on the circle, and L_H is the length-scale, which is here arbitrarily set equal to $L_H = 250km$. Moreover, two non-Gaussian homogeneous correlation tensors have been also defined. The first one is based on Gaspari and Cohn (1999 Eq 4.10) as $C_h(x,y) = \rho_L(x-y)$ with

$$\rho_L(r) = \begin{cases} -\frac{1}{4}\left(\frac{r}{L}\right)^5 + \frac{1}{2}\left(\frac{r}{L}\right)^4 + \\ \quad \frac{5}{8}\left(\frac{r}{L}\right)^3 - \frac{5}{3}\left(\frac{r}{L}\right)^2 + 1, 0 \le r \le L, \\ \frac{1}{12}\left(\frac{r}{L}\right)^5 - \frac{1}{2}\left(\frac{r}{L}\right)^4 + \frac{5}{8}\left(\frac{r}{L}\right)^3 + \frac{5}{3}\left(\frac{r}{L}\right)^2 \\ \quad - 5\left(\frac{r}{L}\right) + 4 - \frac{2}{3}\left(\frac{L}{r}\right), L \le r \le 2L, \\ 0, 2L \le r, \end{cases}$$

and $L = \sqrt{0.3}L_H$ in order to obtain the same theoretical length-scale as in the Gaussian case. The second non-Gaussian homogeneous tensor is similarly defined with the Second Order Auto Regressive (SOAR) correlation (Daley, 1991 p117) $\rho(r) = (1 + \frac{r}{L_H})e^{-\frac{r}{L_H}}$. The spectra on the circle of these three correlations are represented in figure 2.

Then, a heterogeneous correlation is computed using a c-stretching Schmidt transformation (Courtier and Geleyn 1988), adapted to the circle and defined by $h(x) = a\left[\pi - 2Atan\left(\frac{1}{c}tan(\frac{\pi}{2} - \frac{1}{2}\frac{x}{a})\right)\right]$ with $c = 2.4$ (the Schmidt transformation is used for a different purpose in the Arpège global stretched model to obtain a variable resolution). Its inverse is denoted by h^{-1}. The resulting heterogeneous correlation tensor is

$$B(x,y) = B_h\left(h^{-1}(x), h^{-1}(y)\right), \tag{7}$$

which provides correlation functions that are relatively sharp around $180\,^\circ$, and broad around $0\,^\circ$.

A discretized version of these correlation tensors on a given grid leads to covariance matrices, that depend on the resolution of the grid. For a given truncation T, the number of grid points is $N_g = 2T + 1$ and the homogeneous associated resolution is then $\delta x = \frac{2\pi a}{N_g}$. In this paper, we will use $T = 120$ as an example. In this experimental framework, an ensemble of generated background errors is constructed following the method described by Fisher and Courtier (1995) : $\varepsilon_b = \boldsymbol{B}^{1/2}\boldsymbol{\zeta}$, where $\boldsymbol{\zeta}$ is a Gaussian random realization with covariance matrix $\boldsymbol{I}$ and mean equal to zero.

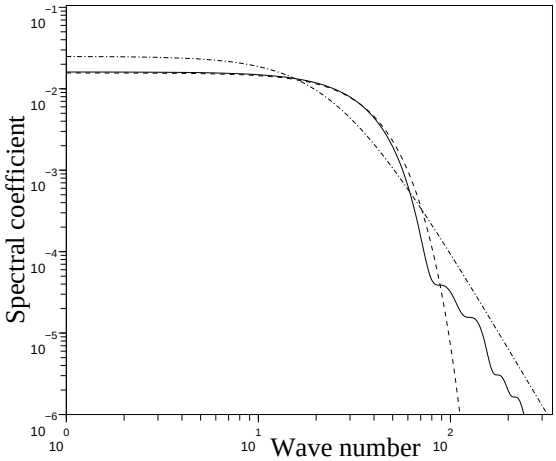


Figure 2. Spectrum of the Gaspari and Cohn correlation (solid line), of the Gaussian correlation (dashed line), and of a SOAR correlation (dash-dotted line).

3.2 Computation of length-scales in a heterogeneous case

For this numerical test, the term $(\sigma(\partial_x \varepsilon_b))^2$ that appears in the B&B length-scale is formally computed as follows. At a point index i, the term is

$$(\sigma(\partial_x \varepsilon_b)_i)^2 = \boldsymbol{\delta}_i^T \boldsymbol{D}\boldsymbol{B}\boldsymbol{D}^*\boldsymbol{\delta}_i,$$

with $\boldsymbol{D}$ the differential operator constructed in Fourier space, and $\boldsymbol{\delta}_i$ the Dirac vector whose value is set to one at index i and set to zero otherwise. In a similar way, the Daley length-scale is computed by a direct computation of the Laplacian at the origin for each correlation function. The Laplacian is computed in Fourier space.

In the 1D framework, the various formulations of length-scale are represented in figure 3 for the heterogeneous G&C-based correlation tensor and for the heterogeneous SOAR-based tensor. The parabola-based and Gaussian-based length-scales are computed as the mean

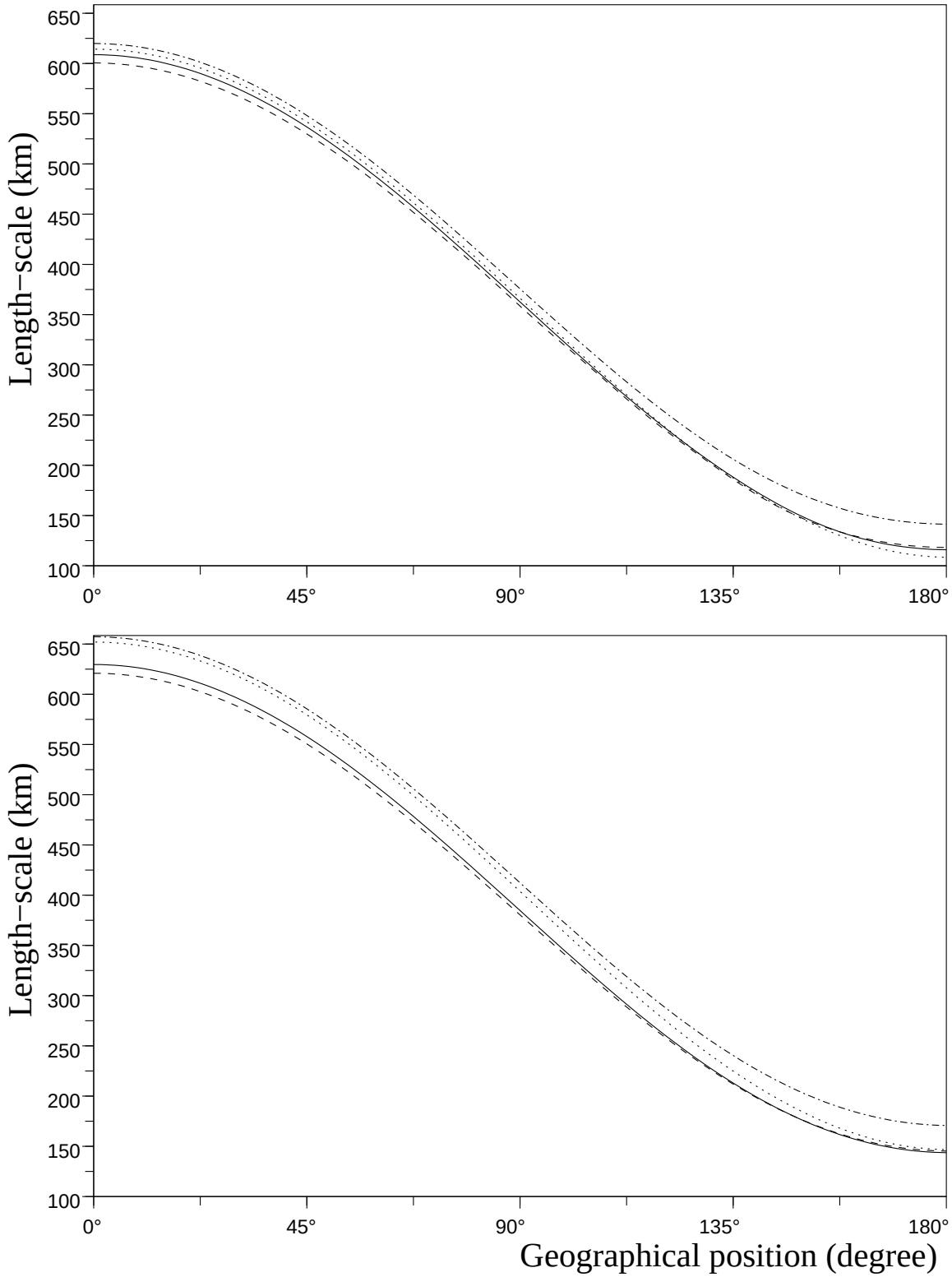


Figure 3. Local length-scales, computed with different formulae, for both G&C correlation tensor (top panel) and SOAR correlation tensor (bottom panel), discretized on the $T120$ circle. Daley (solid line), Belo Pereira-Berre (dashed line), mean parabola-based length-scales (dash-dotted line) and mean Gaussian-based length-scales (dotted line).

value $\frac{1}{2}(L^+ + L^-)$. The Daley length-scale is considered as being the numerical truth and thus the reference.

In the first experiment, represented in the top panel of Fig. 3, the analytical tensor is the heterogeneous G&C-based one. It appears that each length-scale formula is able to represent the geographical variations of the correlation structure : large length-scales near $0\,^\circ$, and small ones near

DOI: 10.1002/qj

180 °. The differences between the various formulations are small. The largest discrepancy is encountered for the parabola-based length-scale.

In the second experiment, represented in the bottom panel of Fig. 3, the analytical tensor is the heterogeneous SOAR-based one. Again, all formulations lead to similar realistic length-scale values.

In the two experiments of Fig. 3 , the parabola-based results are somewhat less accurate than the Gaussian-based length-scales. Moreover, in the bottom panel of Figure 2, the B&B formula is more accurate than the Gaussian-based formula. This indicates that the length-scale diagnosis is slightly sensitive to the underlying correlation function approximation (as mentioned in section 2.3).

Finally, it can be concluded that all formulations lead to similar length-scale values, and the geographical variations are thus well represented in these simulations.

4 Length-scale sampling statistics

4.1 Ensemble size effects in the circle framework

In practical applications, length-scales are usually estimated from a finite ensemble (*e.g.* Belo Pereira and Berre 2006). Figure 4 represents the sampling effect on the estimation of the length-scale L_D for small ensembles. In this experiment, the correlation tensor is the heterogeneous Gaussian tensor on the circle. The true length-scale L_D (dashed line) is compared to estimated length-scales (thin solid line) from 10 members (top) and from 30 members (bottom). Length-scale variations are noised by high frequency variations.

Each sample of N members leads to a particular field of length-scale estimates L^N, which can be considered as a set of random variables. It is thus interesting to know the expectation $\mathbb{E}(L^N)$ of these random variables, and their other statistical characteristics (standard deviation $\sigma(L^N)$, sampling distribution, *etc*).

The expectation function $\mathbb{E}$ is introduced as follows. It is numericaly defined, for a field α, as $\mathbb{E}(\alpha) \approx \frac{1}{N_s}\sum_k \alpha_k$ where α_k are N_s independent realizations of α. For instance, if L^N denotes the length-scale estimated from N members, $\mathbb{E}(L^N) \approx \frac{1}{N_s}\sum_k L^N_k$, where L^N_k is the k^{th} length-scale map estimated from the k^{th} sample of N members. Thereafter, N_s is large and is arbitrarly fixed in order to ensure stable statistics.

Figure 4 shows that the estimated length-scale $\mathbb{E}(L^N)$ is biased : for 10 members, $\mathbb{E}(L^{10}) \neq L$, and the length-scale bias near 0 ° is around $75km$ (12% of total). Moreover, the standard deviation illustrates an even larger distortion of the estimated length-scale, namely by 40% for 10 members (resp. 20% for 30 members).

In order to better understand how the finite size of the ensemble influences the estimation, the sampling distribution of the length-scale can be computed experimentally. In the particular case of Parabola-based and Gaussian-based length-scales, the sampling distribution can also be deduced analytically, from the sampling distribution of the correlation ρ^N between two points separated by a distance δx. This is shown in the appendix.

4.2 Gaussian-based length-scale sampling distribution

The experimental frequency distribution is represented in figure 5 for $N = 25$. It shows that the sampling distribution is positively skewed, and the existence of a bias $b^N_{Gb} = \mathbb{E}(L^N_{Gb}) - L^\infty_{Gb}$. These experimental results are consistent with analytical studies, as shown in the appendix. Positive skewness implies that large length-scale values are often encountered with such ensemble sizes.

The full line in Fig. 6 represents the relative error percentage associated to the bias $\frac{\mathbb{E}(L^N_{Gb})-L^\infty_{Gb}}{L^\infty_{Gb}}$ for a given discretization δx. In that case, the $T120$ discretized circle is considered ($\delta x \approx 166km$) and $L^\infty_{Gb} = 250km$. The error is large for a small ensemble and tiny for a large one. The convergence to the infinite-ensemble value is relatively fast, as it is in $\mathcal{O}(N^{-1})$. For 10 members, the bias ratio is 10%.

However the standard deviation is larger as shown now. Figure 6 shows the ratio $\frac{\sigma^N_{L_{Gb}}}{L^\infty_{Gb}}$ where $\sigma^N_{L_{Gb}} = \sqrt{\mathbb{E}\left\{\left(L^N_{Gb} - \mathbb{E}(L^N_{Gb})\right)^2\right\}}$. The behaviour in $\mathcal{O}(N^{-1/2})$ is observed as expected (see the appendix). For 10 members, the ratio is 40%. This illustrates the predominance of the error standard deviation over the error bias in the length-scale estimation.

4.3 Comparison with other length-scale formulae

Trying to find the sampling distribution of Daley or B&B length-scale analytically is not easy, because it depends on the shape of the function and not only on one correlation. However, numerical experiments indicate a similar behavior to that of the Pb and Gb cases. Figure 7 shows the sample distribution of length-scale for the four formulations: Daley, B&B, Pb and Gb. These length-scales are estimated from a 10 member ensemble. The true correlation tensor used here is a homogeneous Gaussian correlation tensor over the $T120$ discetized circle with length-scale L_H. It appears that the sampling distributions are similar to each other. In particular, both Daley and B&B length-scales present some bias.

4.4 The spatial structure of sampling noise

As illustrated in Fig. 4, the spatial variations of the estimated length-scales tend to be random and spatially uncorrelated, compared to variations of the exact length-scales. This suggests that the estimated length-scale field is affected by a sampling noise, whose amplitude is relatively large in the small scales (compared to the exact length-scale field).

In order to explore this issue, energy spectra have been calculated for geographical maps of the exact length-scales, of the ensemble-estimated length-scales and of the corresponding estimation errors. The results are shown in Fig. 8. As expected from Fig. 4, the ensemble-estimated

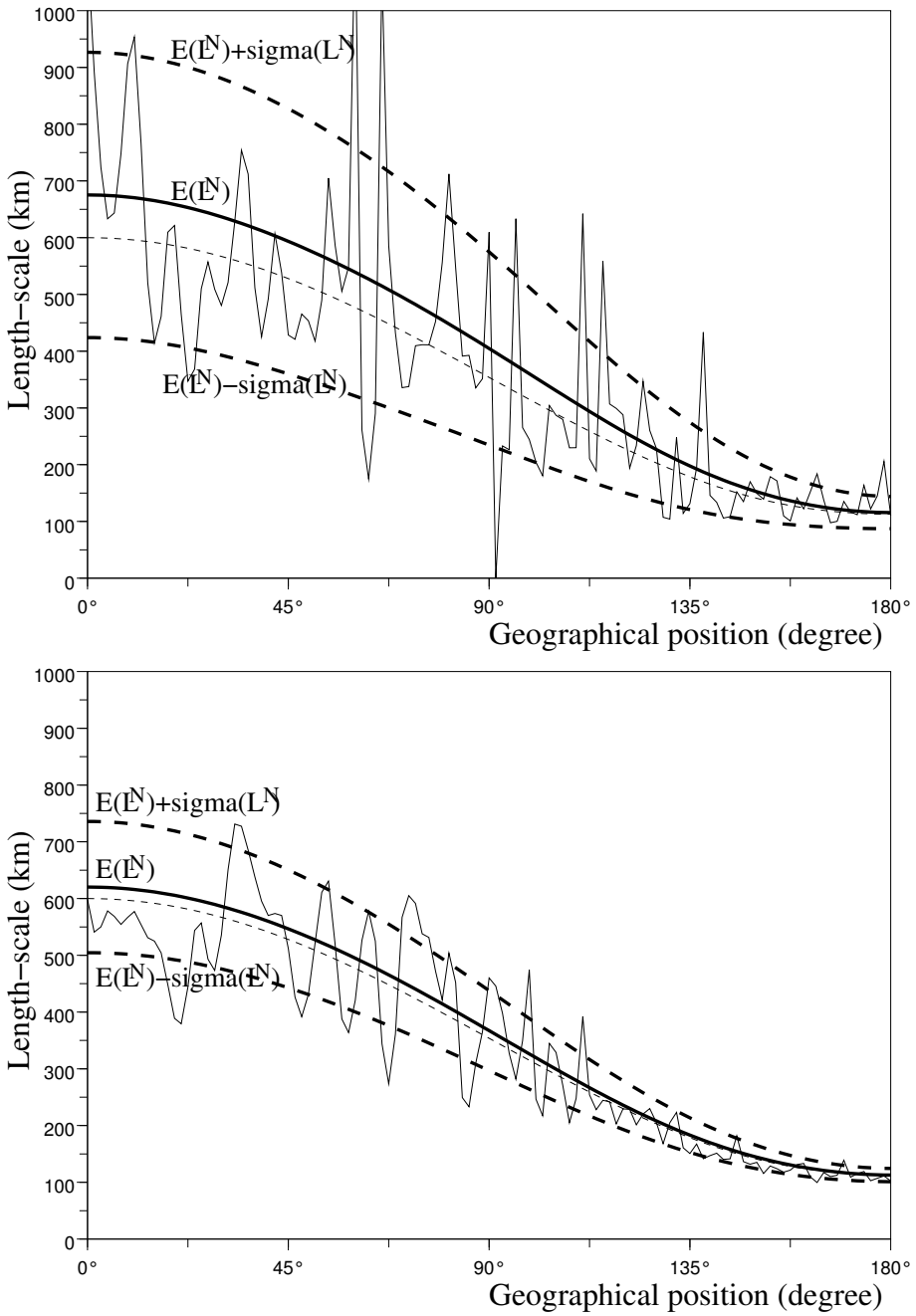


Figure 4. Sensitivity of length-scale estimation to the ensemble size. The true length-scale map (thin dashed line) is compared with the estimated length-scale (thin solid line) for $N = 10$ members (top panel) and $N = 30$ members (bottom panel). The curve of the expectation $\mathbb{E}(L^N)$ (bold solid line) illustrates the existence of a bias, and the usual range $\mathbb{E}(L^N) \pm \sigma(L^N)$ (bold dashed lines, where $\sigma(L^N)$ is the standard deviation of L^N) offers a representation of the expected range of values reached by the estimated length-scales.

length-scale maps spuriously contain much more small scale energy than the exact length-scale map. This corresponds to the artificial contribution of sampling noise, whose energy spectrum is close to a white noise.

These results indicate that spatial filtering techniques based on spectral or wavelet techniques may be worth considering. This is illustrated in the next subsection.

4.5 Sampling noise reduction through spatial filtering

Background error correlation modeling is often based on a spectral diagonal approach (Courtier *et al.*, 1998). More recently, Fisher (2003) has also defined an error correlation modeling with a wavelet diagonal approach. As discussed in Pannekoucke *et al.* (2007), using these techniques amounts to spatially averaging the local correlation functions.

In the spectral diagonal approach, this spatial averaging is global, in the sense that it is calculated as a uniform

DOI: 10.1002/qj

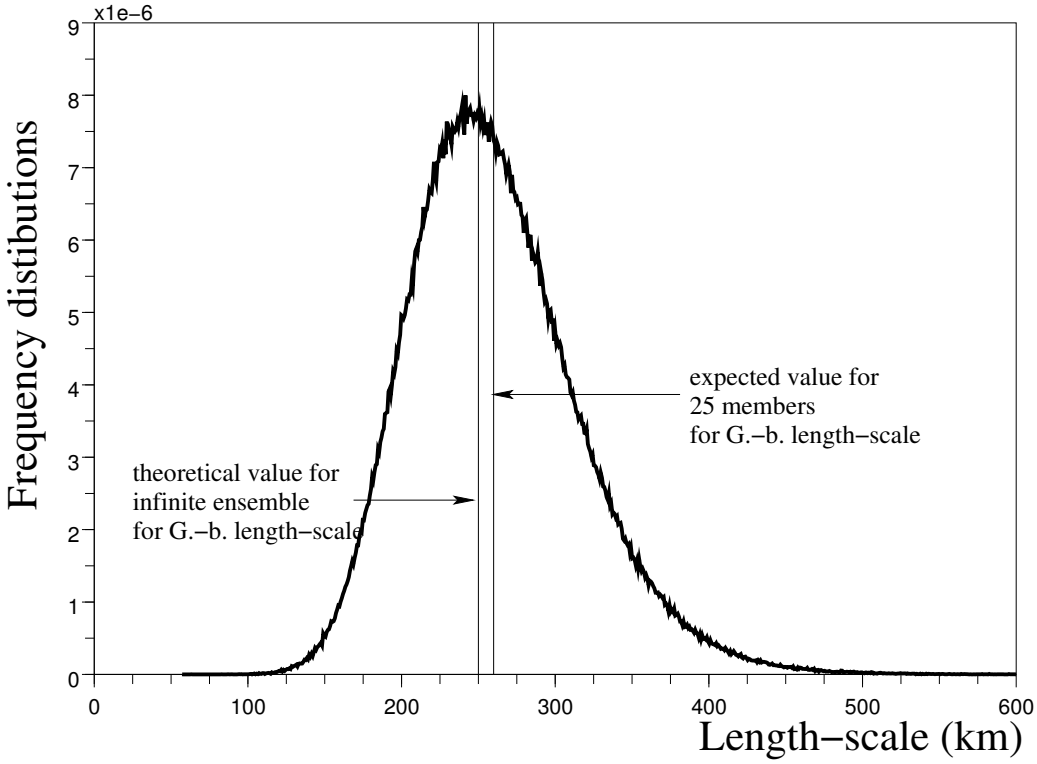


Figure 5. Sampling distribution of the Gaussian-based length-scale for $\delta x = 166\,km$ and 25 members. The theoretical length-scale $L_H = 250\,km$ is over-estimated by the expected length-scale from the finite ensemble.

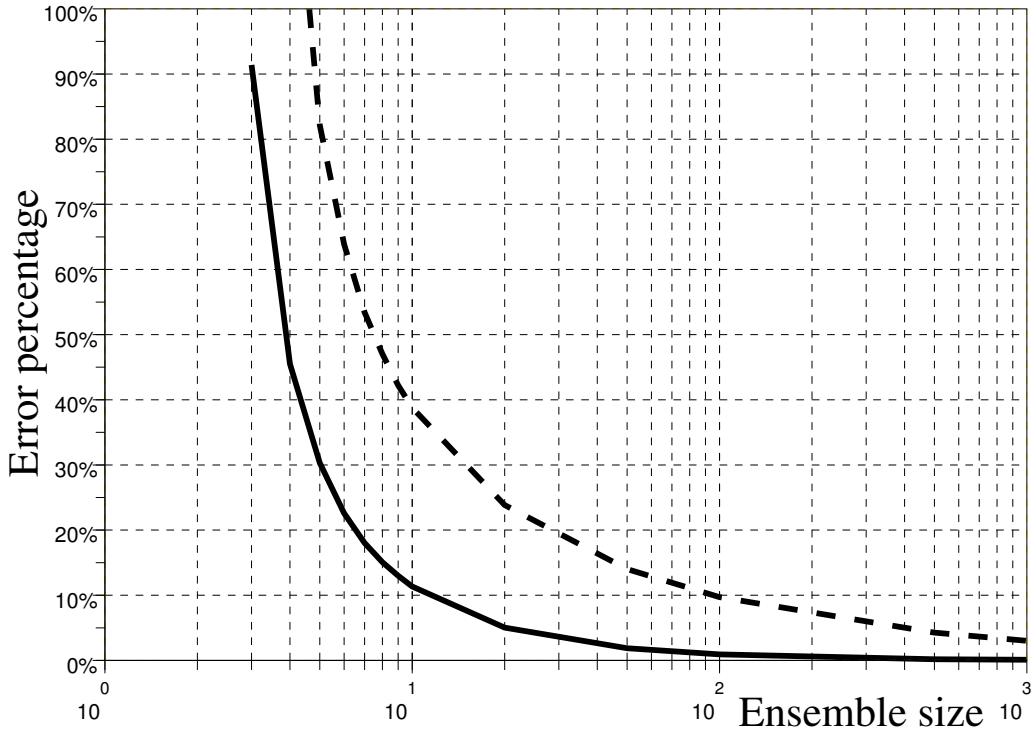


Figure 6. Convergence of length-scale error with the ensemble size: the bias (solid line) and the standard deviation (dashed line), both normalized by the true length-scale. (See text for details.)

average over the whole domain. In the wavelet diagonal approach, this spatial averaging is rather local. This means that wavelets allow the size of the statistical sample to be increased, by introducing a local spatial sample (multiplied by the ensemble sample), while keeping the

DOI: 10.1002/qj

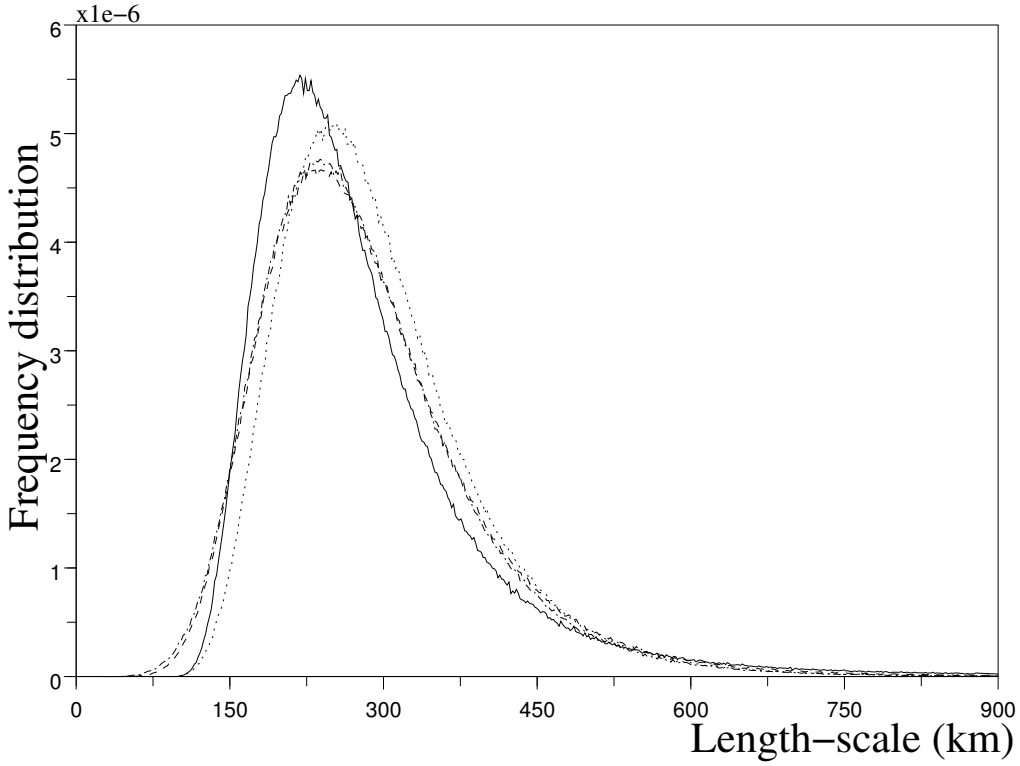


Figure 7. Comparison of sampling distributions of the length-scale, estimated from a 10 member ensemble. The correlation tensor is set equal to the Gaussian homogeneous tensor associated to the length-scale $L_H = 250\,km$ and truncation $T = 120$. Estimated length-scales : Daley (solid line), Gaussian-based (dash-doted line), parabola-based (dotted line) and Belo Pereira& Berre (dashed line). Note that Gb is almost superposed with B&B.

possbility to represent geographical variations.

The efficiency of this wavelet filtering approach has been illustrated by Pannekoucke *et al.* (2007) in a 1D heterogeneous case. Here, we will focus on a 1D homogeneous case, in order to illustrate the effect of spatial filtering on the bias and standard deviation of the length-scale error.

This 1D case corresponds to a homogeneous Gaussian correlation tensor, discretized on a T120 circle (associated to $N_g = 241$ grid points), with a theoretical length-scale equal to $L_H = 250km$. An estimation of the correlation matrix, with an ensemble of N members, leads to a heterogeneous covariance matrix. At a given point k, the Gaussian-based length-scale at this point is the mean length-scale $L_e^N = (L^+ + L^-)/2$. The correlation modelled with the diagonal assumption in spectral space is homogeneous, and corresponds to the average of the N_g estimated correlation functions. The resulting Gaussian based length-scale is thus $L_{ds} = \left(L(\bar{\rho}^+) + L(\bar{\rho}^-)\right)/2$ where $\bar{\rho}^+ = \frac{1}{N_g}\sum_k \rho_k^+$ and $\bar{\rho}^- = \frac{1}{N_g}\sum_k \rho_k^-$.

Different random variables are also introduced: L_{ds}^N is the length-scale resulting from a spectral diagonal assumption estimated with an ensemble of N members ; L_{dw}^N is the corresponding length-scale resulting from a wavelet diagonal assumption. Relative errors $\mathbb{E}(L_e^N)/L_H - 1$ (solid line), $\mathbb{E}(L_{ds}^N)/L_H - 1$ (dash-dotted line) and $\mathbb{E}(L_{dw}^N)/L_H - 1$ (dash line) are represented on the top panel of Fig. 9 for ensemble size $N \in [6, 200]$. This error corresponds to the bias normalized by the length-scale L_H. As shown in section 4.2, $\mathbb{E}(L_e^N)/L_H - 1$ converges as $\mathcal{O}(N^{-1})$, while $\mathbb{E}(L_{ds}^N)/L_H - 1$ is close to zero everywhere. $\mathbb{E}(L_{dw}^N)/L_H - 1$ is small, although it remains different from zero even for a large ensemble. This is due to a known defect of the wavelet diagonal assumption : length-scale can be under or over-estimated with as much as 10% error (Pannekoucke *et al.*, 2007).

Then to appreciate the accuracy of the estimation, the ratios σ_e^N/L_H, σ_{ds}^N/L_H and σ_{dw}^N/L_H are also represented (bottom panel of Fig. 9). These ratios represent the error standard deviation normalized by the length-scale L_H. Again, σ_e^N/L_H converges to zero as $\mathcal{O}(N^{-1/2})$. σ_{ds}^N/L_H converges at the same rate but with a factor close to $1/17 \approx 1/\sqrt{N_g}$: for $N = 6$, $\sigma_e^N/L_H \approx 50\%$, while $\sigma_{ds}^N/L_H \approx 3\%$. The convergence of σ_{dw}^N/L_H is between these two extreme convergences : it has the same rate as the ensemble, but with a factor close to $1/5$: for $N = 6$, $\sigma_e^N/L_H \approx 50\%$ while $\sigma_{dw}^N/L_H \approx 10\%$.

These results illustrate the property of *e.g.* a wavelet formulation to represent the length-scale values with a better accuracy (here a factor 5) than the direct ensemble estimation.

DOI: 10.1002/qj

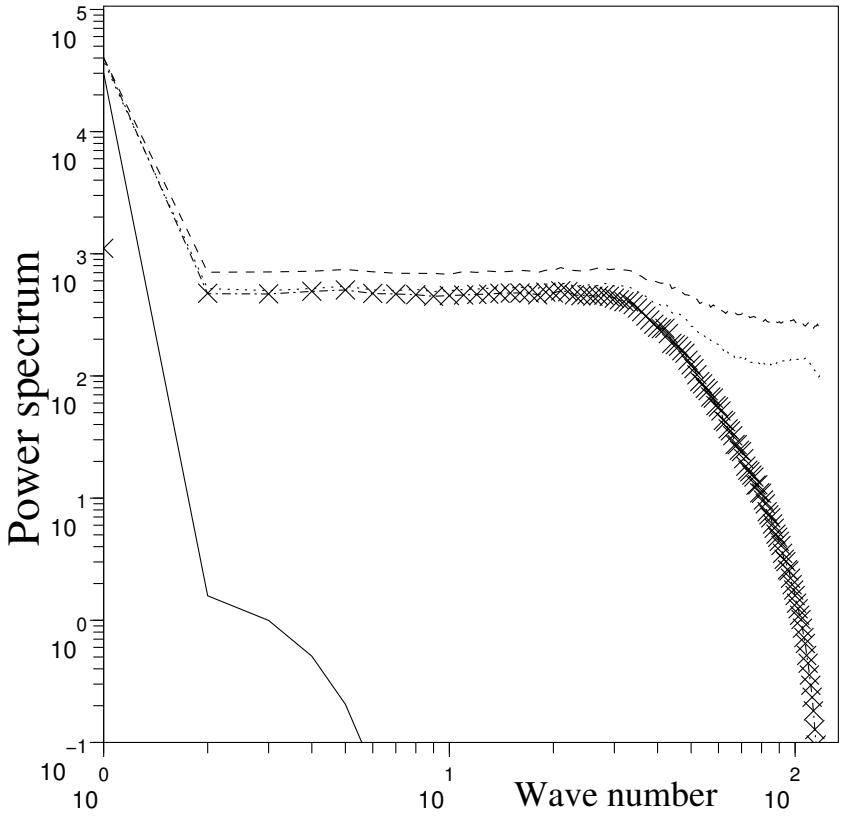


Figure 8. Energy spectra for geographical maps of the exact length-scales (solid line) and of the ensemble estimated length-scales, with 10 members: Daley (dashed line), B&B (dotted line) and Gb (dash-dotted line). Pb (not shown) is similar to Gb. The energy spectrum of the Gb estimation errors is represented by crosses, which are almost superposed with the Gb map energy, except for wavenumber 1. Note also that the spectrum of the exact length-scales (solid line) reflects the predominance of wave number 1, in accordance with figure (3).

5 Application to an ensemble of NWP forecasts

An application to an ensemble of NWP forecasts has been studied, by using an operational non-stretched version of the Arpège model (Courtier and Geleyn, 1988), whose assimilation system is a 4D-Var scheme (Rabier *et al.*, 2000; Veersé and Thépaut, 1998). The background error covariance matrix is calculated by using an ensemble of perturbed assimilation runs (Houtekamer *et al.*, 1996, Fisher 2003). The detailed results for this Arpège ensemble are described in Belo Pereira and Berre (2006).

The available ensemble consists in a set of 6 forecast differences for each day of the period 9 February to 24 March 2002, and time-averaged covariances are calculated over this 49-day period. Figure 10 presents the results obtained with the B&B zonal length-scale (top panel) and with the Gaussian-based zonal length-scale (bottom panel) for the logarithm of surface pressure. As in the previous subsection, the zonal gradient in the B&B length-scale is computed in spectral space. Each formulation represents well the land-sea contrast, and the influence of the orography, with *e.g.* larger values over tropical oceans and smaller values near the Andes. Actually there are only slight differences between the two formulations of length-scale. This supports the idea that the Gaussian-shape assumption near the origin is acceptable, leading to realistic length-scale values.

It may be mentioned that such maps of length-scales provide a full vision of geographical variations in the curvature of correlation functions. Such geographical variations can thus be examined with more details than when only plotting correlation functions at a few selected points on the globe (as *e.g.* in Baker *et al.*(1987)). On the other hand, it should be reminded that length-scales give information about the correlation curvature near the origin only, while full correlation functions provide information about all separation distances. These two diagnostics are thus to be seen as complementary.

6 Conclusion

Some approximations of the theoretical Daley (1991, p110) length-scale have been discussed in this paper. In particular, an economical estimation based on a Gaussian assumption has been investigated. Firstly, it has been shown in a 1D heterogeneous context that the different length-scale formulae provide similar realistic length-scale values and variations.

Secondly, a study of the sampling distribution of the estimated length-scales has been carried out, both analytically and experimentally. It has been shown that the estimated length-scales are affected by a positive bias when the ensemble size N is small. This bias converges towards zero in $\mathcal{O}(N^{-1})$. This bias has been shown to be

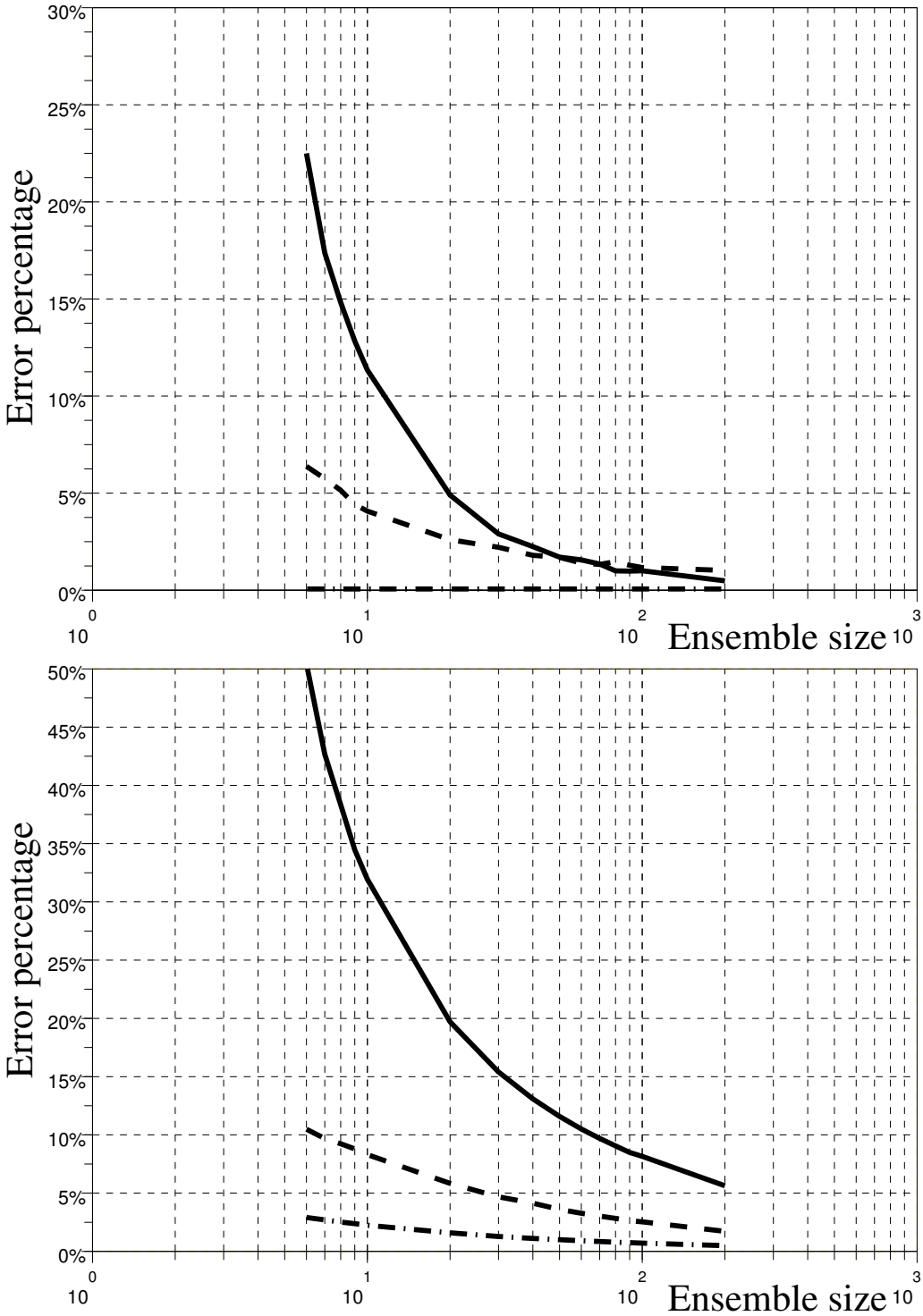


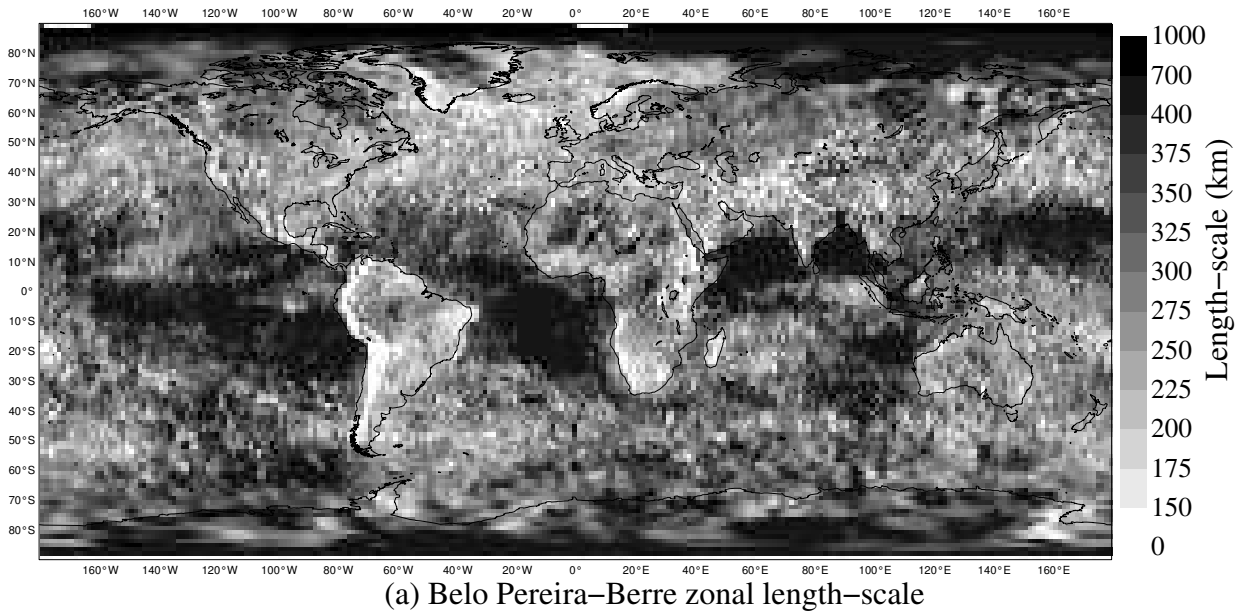
Figure 9. Comparison of the convergence of length-scale error with the ensemble size : directly estimated (solid line), resulting from a diagonal asumption in wavelet space (dashed line) and in spectral space (dash-dotted line). Top panel : bias normalized by L_H. Bottom panel : standard deviation normalized by L_H. (See text for details.)

smaller than the estimation error standard deviation. The latter converges towards zero in $\mathcal{O}(N^{-1/2})$.

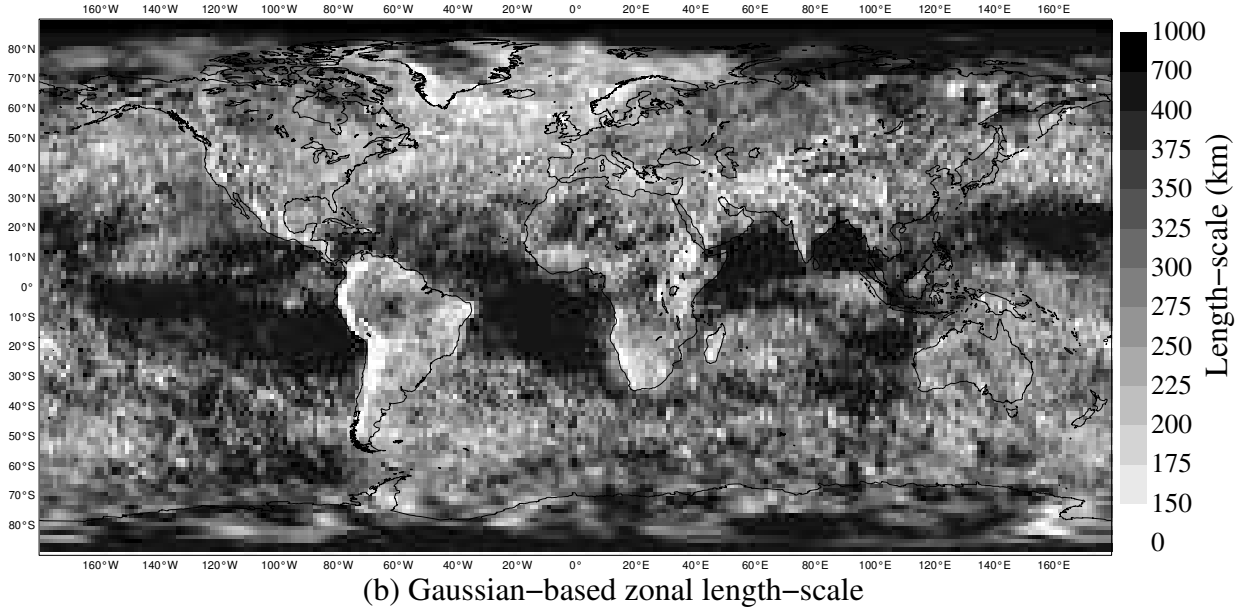
In addition, the examination of length-scale geographical variations and of their energy spectrum indicates that the sampling noise tends to be uncorrelated spatially (typically like white noise). This suggest that local space averaging techniques, such as those based on wavelets, are worth considering in order to spatially filter sampling noise.

Finally, the Belo Pereira and Berre formula has been compared to the Gaussian-Based length-scale on a 2D spherical example from a NWP ensemble data set. Length-scale values and variations appear to be similar according to the two formulae. This indicates that the

DOI: 10.1002/qj



(a) Belo Pereira–Berre zonal length–scale



(b) Gaussian–based zonal length–scale

Figure 10. Zonal length-scale of the surface pressure logarithm, numerically computed with Belo Pereira-Berre (a) and Gaussian-based (b) formulae. This is a 'climatological' average, calculated over 49 days.

assumption that the shape of the correlation function is Gaussian is reasonable in order to estimate the length-scale (defined by Daley (1991)).

The possibility to calibrate a correlation model from length-scale estimates is another potential application of the considered formulae in this paper, even if this issue has not been investiagated here. A relatively obvious limitation is that the knowledge of the length-scales may not be sufficient to determine an accurate model of the whole correlation functions, e.g. because length-scales characterize the curvature of correlation functions near their origin only. With this perspective in mind, correlation modelling based *e.g.* on wavelets may be more attractive than modelling based *e.g.* on a Gaussian approximation and on a specification of length-scales only.

7 Appendix

7.1 Approximation of the correlation sampling distribution

The Normal distribution of a correlated background error pair $\boldsymbol{\varepsilon}_b = (\varepsilon_1, \varepsilon_2)$ may be written as $f_b(\boldsymbol{\varepsilon}_b) = \frac{1}{2\pi|\boldsymbol{B}|^{1/2}} exp\left(-1/2||\boldsymbol{\varepsilon}_b||^2_{\boldsymbol{B}^{-1}}\right)$, where the covariance matrix is $\boldsymbol{B} = \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}$, whose determinant is $|\boldsymbol{B}|$ and whose correlation is ρ. From N sample values $(\varepsilon_1^1, \varepsilon_2^1) \ldots (\varepsilon_1^N, \varepsilon_1^N)$, the corresponding sample variances are $S_1^2 = \frac{1}{N}\sum_k {\varepsilon_1^k}^2$, $S_2^2 = \frac{1}{N}\sum_k {\varepsilon_2^k}^2$, and the sample correlation is $C = \frac{1}{N}\sum_k \varepsilon_1^k\varepsilon_2^k/(S_1S_2)$. Of course, S_1^2, S_2^2 and C are random variables.

Fisher (1953) has expressed the sampling distribution of C (Kendall *et al.*, 1998, Hotteling, 1953). But this formulation is in fact too complex, and some approximations of this sampling distribution must be used. For a large ensemble, the distribution is close to Gaussian. For a small ensemble, the distribution is not Gaussian, and its skewness increases with the correlation value. When N is not too small, typically $N \geq 25$ members, Fisher has proposed a suitable transformation, where the convergence to Gaussianity of the new variable is accelerated. Then the random variable $Z = tanh^{-1}C$ follows, with a good approximation, a Gaussian distribution:

$$Z \sim \mathcal{N}\left(\mu_Z(\rho,N), \sigma_Z^2(\rho,N)\right), \tag{8}$$

with the mean $\mu_Z(\rho,N) = \zeta + \frac{\rho}{2(N-1)} + \frac{\rho(5+\rho^2)}{8(N-1)^2} + \frac{\rho(11+2\rho^2+3\rho^4)}{16(N-1)^3} + \mathcal{O}\left(N^{-4}\right)$ where $\zeta = \frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right)$ and the standard deviation $\sigma_Z(\rho,N)^2 = \frac{1}{N-1} + \frac{4-\rho^2}{2(N-1)^2} + \frac{22-6\rho^2-3\rho^4}{6(N-1)^3} + \mathcal{O}\left(N^{-4}\right)$. A simple change of variable leads to the sampling distribution of correlation

$$f_C(c) = \frac{1}{(1-c^2)\sigma_Z(\rho,N)\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\frac{\left[Z(c)-\mu_Z(\rho,N)\right]^2}{\sigma_Z(\rho,N)^2}\right\}, \tag{9}$$

where by definition $P(C \in [c, c+dc]) = f_C(c)dc$, where P is the probability measure. The top panel of figure 11 shows this sampling distribution (bold solid line) for $N = 25$ members and $\rho = exp(-\delta x^2/2L_H^2) \approx 0.8$. An experimental frequency distribution (solid line) is obtained numerically for $N = 25$. It illustrates the accuracy of the Fisher approximation.

As suggested by the above equations of μ_Z and of σ_Z^2, the bias and standard deviation of Z converge towards zero in $\mathcal{O}(N^{-1})$ and in $\mathcal{O}(N^{-1/2})$ respectively. The rates of convergence are similar for the correlation C.

7.2 Approximation of the Gaussian-based length-scale sampling distribution

Applying the Fisher's transformation to length-scale leads to the sampling distribution of length-scales. The calculation is given for the Gaussian-based length-scale, knowing that the parabola-based case is similar.

Actually, for the Gb, correlation must be positive. Thus the correlation sampling distribution has to be limited to the positive correlation part. Let $\chi_{(0,1]}$ be the characteristic function defined on $[-1,1]$; it is equal to one on $(0,1]$ and null otherwise. The random variable associated to positive correlation is $C^+ = \chi_{(0,1]}C$. Its sample distribution is $f_{C^+}(c) = \Lambda(\rho,N)^{-1} f_C(c)$, $c > 0$, where $\Lambda(\rho,N) = \int_0^1 f_C(c)dc$ is the normalization term ; it can be approximated by $\Lambda(\rho,N) \approx 1 - \frac{\Gamma(N)}{\Gamma(N+1/2)\sqrt{2\pi}}\frac{(1-\rho^2)^{N/2}}{\rho}$ (Hotteling, 1953). The change of variable $C^+ = \exp\left(-\frac{\delta x^2}{2L_{Gb}^{N\ 2}}\right)$, with L_{Gb}^N the estimator of the Gb length-scale calculated with a N member ensemble, leads to the sampling distribution

$$f_{L_{Gb}^N}(l) = \frac{\Lambda(\rho,N)^{-1}\delta x^2}{2l^3 \sinh(\frac{\delta x^2}{2l^2})\sigma_Z(\rho,N)\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\frac{\left[Z\left(\rho(l)\right)-\mu_Z(\rho,N)\right]^2}{\sigma_Z(\rho,N)^2}\right\}, \tag{10}$$

with $\rho(l) = \exp\left(-\frac{\delta x^2}{l^2}\right)$. The bottom panel of figure 11 illustrates the frequency distribution and approximation of the sampling distribution for both Pb and Gb length-scales. These results are obtained with the correlation value $\rho \approx 0.8$ of the previous section. It appears that the analytical approximations of sampling distribution are in accordance with the experimental frequency distribution, with a sufficient accuracy.

In the case of an infinite ensemble, the expected value is $L_{Gb}^{\infty} = \frac{\delta x}{\sqrt{-2\,ln(\rho))}}$. In the N-member case, the resulting sampling distribution is positively skewed, and L_{Gb}^N is a positively biased estimator: $b_{Gb}^N = \mathbb{E}(L_{Gb}^N) - L_{Gb}^{\infty} > 0$. Again it can be deduced from the analytical approximation of the sampling distribution that, when N is large there is a convergence to zero in $\mathcal{O}(N^{-1})$ for the bias, and in $\mathcal{O}(N^{-1/2})$ for the standard deviation.

This result is also valid for other length-scale formulae under the assumption that background error distribution is Gaussian.

References

Baker W., Bloom S., Woollen J., Nestler M., Brin E., Schlatter T. and Branstator G. 1987. *Experiments with a three-dimensional statistical objective analysis scheme using FGGE data. Mon. Wea. Rev.*, **115**, 272–296.

Belo Pereira M. and Berre L. 2006. *The use of an Ensemble approach to study the Background Error Covariances in a Global NWP model. Mon. Wea. Rev.*, **134**, 2466–2489.

Bouttier F. 1993. *The dynamics of error covariances in a barotropic model. Tellus*, **45A**, 408–423.

Buehner M. 2005. *Ensemble-derived stationary and flow-dependent background-error covariances: Evaluation in a quasi-operational NWP setting. Q.J.R. Meteorol. Soc.*, **131**, 1013–1043.

Courtier P. and Geleyn J.F. 1988. *A global numerical weather prediction model with variable resolution : Application to the shallow-water equations. Q.J.R. Meteorol. Soc.*, **114**, 1321–1346.

Courtier P., Andersson E., Heckley W., Pailleux J., Vasiljević D., Hamrud M., Hollingsworth A., Rabier F. and Fisher M. 1998. *The ECMWF implementation of three-dimensional variational assimilation (3D-Var). I: Formulation. Quart. J. Roy. Meteor. Soc.*, **124**, 1783–1807.

Daley R. 1991. *Atmospheric Data Analysis*. Cambridge University Press. p471.

Deckmyn A. and Berre L. 2005. *A wavelet approach to representing background error covariances in a LAM. Mon. Wea. Rev.*, **133**, 1279-1294.

Fisher R.A. 1953. *On the 'probable error' of a coefficient of correlation deduced from a small sample*. Metron, **1**, 1–32.

Fisher M. and Courtier P. 1995. *Estimating the covariance matrices of analysis and forecast error in variational data assimilation.* ECMWF Technical Memorandum, **220**, 29pp.

Fisher M. 2003. *Background error covariance modelling*. Processing of the ECMWF Seminar on "Recent developments in data assimilation for atmosphere and ocean", Reading, 8–12 September 2003, 45–63.

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

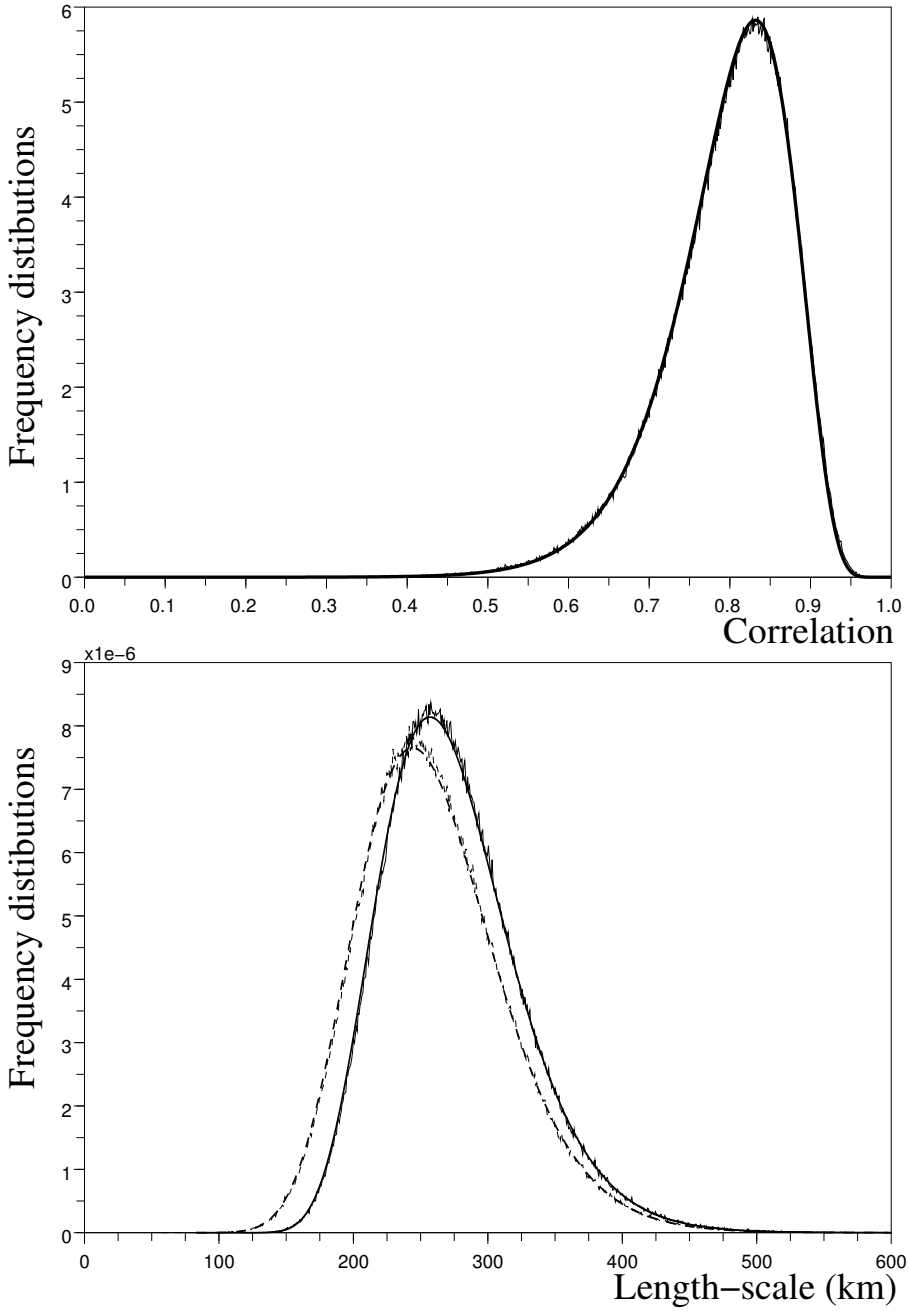


Figure 11. Comparison of approximated and estimated sampling distributions resulting from a 25 member ensemble. Top panel : sampling distribution for correlation $\rho \approx 0.8$, either theoretically approximated (bold solid line), or estimated experimentally (thin solid line). Bottom panel : sampling distribution of both parabola-based and Gaussian-based length-scales for $\delta x = 166\,km$. The theoretical approximation of the parabola-based length-scale (resp. Gaussian-based) is in bold solid line (resp. bold dashed line), while its sampled estimation is in thin solid line (resp. thin dashed line).

Hollingsworth A. 1987. *Short- and medium-range numerical weather prediction*. Collection of papers presented at the WMO/IUGG symposium, Tokyo, 4–8 August 1986.

Houtekamer P.L., Lefaivre L., Derome J., Ritchie H. and Mitchell H.L. 1996. *A system simulation approach to ensemble prediction*. *Mon. Wea. Rev.*, **124**, 1225–1242.

Houtekamer P.L. and Mitchell H.L. 2001. *A sequential ensemble Kalman filter for Atmospheric Data Assimilation*. *Mon. Wea. Rev.*, **129**, 123–137.

Hoteling H. 1953. *New light on the correlation coefficient and its transforms*. *Journal of the Royal Statistical Society*. Series B (Methodological), **15**, 193–232.

Ingleby B. 2001. *The statistical structure of forecast errors and its representation in The Met. Office Global Model*. *Q.J.R. Meteorol. Soc.*, **124**, 1783–1807.

Kalnay E. 2002. *Atmospheric modeling, data assimilation and predictability*. Cambridge University Press, p364.

Kendall M., Stuart A. and Ord J.K. 1998. *Kendall's Advanced Theory of Statistics, Volume 1: Distribution Theory*. A Hodder Arnold Publication.

Pannekoucke O., Berre L. and Desroziers G. 2007. *Filtering properties of wavelets for local background-error correlations*. *Q.J.R. Meteorol. Soc.*, **133**, 363–379.

Rabier F., McNally A., Andersson E., Courtier P., Undén P., Eyre J., Hollingsworth A. and Bouttier F. 1998. *The ECMWF implementation of three-dimensional variational assimilation (3D-Var). II: Structure*

functions. *Quart. J. Roy. Meteor. Soc.*, **124**, 1809–1829.

Rabier F., Jarvinen H., Klinker E., Mahfouf J.F. and Simmons A. 2000. *The ECMWF operational implementation of four-dimensional variational assimilation. I: Experimental results with simplified physics*. *Q.J.R. Meteorol. Soc.*, **126**, 1148–1170.

Veersé F. and Thépaut J-N. 1998. *Multiple-truncation incremental approach for four-dimensional variational data assimilation*. *Q.J.R. Meteorol. Soc.*, **124**, 1889–1908.