Ocean Sci., 8, 121–142, 2012
www.ocean-sci.net/8/121/2012/
doi:10.5194/os-8-121-2012

Ocean Science

Towards an improved description of ocean uncertainties: effect of local anamorphic transformations on spatial correlations

J.-M. Brankart[1], C.-E. Testut[2], D. Béal[1], M. Doron[1], C. Fontana[1], M. Meinvielle[1], P. Brasseur[1], and J. Verron[1]

[1]LEGI/CNRS, UMR5519, Grenoble, France
[2]Mercator-Océan, Toulouse, France

Correspondence to: J.-M. Brankart (jean-michel.brankart@hmg.inpg.fr)

Received: 20 September 2011 – Published in Ocean Sci. Discuss.: 28 October 2011
Revised: 23 February 2012 – Accepted: 24 February 2012 – Published: 6 March 2012

Abstract. The objective of this paper is to investigate if the description of ocean uncertainties can be significantly improved by applying a local anamorphic transformation to each model variable, and by making the assumption of joint Gaussianity for the transformed variables, rather than for the original variables. For that purpose, it is first argued that a significant improvement can already be obtained by deriving the local transformations from a simple histogram description of the marginal distributions. Two distinctive advantages of this solution for large size applications are the conciseness and the numerical efficiency of the description. Second, various oceanographic examples are used to evaluate the effect of the resulting piecewise linear local anamorphic transformations on the spatial correlation structure. These examples include (i) stochastic ensemble descriptions of the effect of atmospheric uncertainties on the ocean mixed layer, and of wind uncertainties or parameter uncertainties on the ecosystem, and (ii) non-stochastic ensemble descriptions of forecast uncertainties in current sea ice and ecosystem pre-operational developments. The results indicate that (i) the transformation is accurate enough to faithfully preserve the correlation structure if the joint distribution is already close to Gaussian, and (ii) the transformation has the general tendency of increasing the correlation radius as soon as the spatial dependence between random variables becomes nonlinear, with the important consequence of reducing the number of degrees of freedom in the uncertainties, and thus increasing the benefit that can be expected from a given observation network.

1 Introduction

As a result of inescapable inaccuracies or approximations in the observations and in the models, uncertainties are inherent to any description or simulation of the real ocean. A realistic and efficient modelling of these uncertainties is of key importance for many oceanographic applications: (i) to objectively check simulation results against independent observations, (ii) to optimally assimilate data, and thus obtain the maximum benefit from an expensive, but incomplete, observing system, and (iii) to rationaly design future observation networks. It is thus essential to the production and use of ocean operational data, as delivered for instance by the MyOcean system[1], which is the target application of this study.

Ensemble (or Monte Carlo) methods provide a good way of describing uncertainties in ocean dynamical systems, by explicitly exploring how uncertainties in the governing laws, parameters or forcings (the prior information) propagate to the observed quantities or to the operational products (Palmer et al., 2005; Lermusiaux, 2006). However, even if an explicit stochastic modelling is used to solve a practical problem, there is often a strong temptation (in large size applications) to simplify the result using a Gaussian model, because it is much more efficient (i) to describe the uncertainties (by the mean and covariance), and (ii) to assimilate observations (using linear update formulas, as in the ensemble Kalman filter, see Evensen and van Leeuwen, 1996). Without a prior assumption about the shape of the probability distribution, large size problems are indeed very complex in general (van Leeuwen, 2009, 2010; Bocquet et al., 2010), mainly because the size of the sample that is required to identify a general multivariate distribution increases exponentially with

[1] http://www.myocean.eu.org/

Published by Copernicus Publications on behalf of the European Geosciences Union.

the number of dimensions (curse of dimensionality). To circumvent this difficulty, one possible simplification is to look for univariate nonlinear changes of variables (anamorphosis transformations) transforming the marginal distribution of each random variable into a Gaussian distribution. One-dimensional probability distributions can indeed be identified with a much smaller sample, and it may well happen that such a separate transformation for each random variable also helps improving the Gaussianity of their joint distribution (although this needs to be checked in every practical application). This technique originates from geostatistics (Wackernagel, 2003) and was first introduced in oceanography by Bertino et al. (2003), in the framework of the ensemble Kalman filter.

However, the studies presented in Bertino et al. (2003) and later in Simon and Bertino (2009) were directly focused on the impact that anamorphic transformations may have on the performance of the ensemble Kalman filter, without much emphasis on the improvements in the multivariate statistics. In this context, they also preferred to apply the same transformation over the whole model domain (but different for each model variable), so that a much larger sample is available to identify the transformation function. Yet, if the objective is also to propose a generic method (beyond the Gaussian scheme) to improve the description of the uncertainties, which can be spatially inhomogeneous, any practical possibility of extending this towards local anamorphic transformations should be evaluated. In a recent paper, Béal et al. (2010) proposed a very simple algorithm to obtain such local transformations, and started evaluating its potential for describing a 30-day ensemble forecast of the North-Atlantic ecosystem (simulating the effect of wind uncertainties). However, the paper was exclusively focused on the improvement of local correlations (at given locations) between phytoplankton and the other ecosystem compartments (nutrients, zooplankton), in the perspective of ocean colour data assimilation. Yet, with an algorithm working locally (i.e. transforming each model grid point with a different anamorphosis function), it is also important to study how the spatial correlations are modified, and hopefully improved, by the transformation.

The purpose of the present paper is thus to evaluate the effect of local anamorphic transformations on spatial correlations for various kinds of ocean uncertainties. The study includes, on the one hand, the stochastic ensemble description of the ocean mixed layer response to atmospheric forcing uncertainties (Sect. 3), the ecosystem response to wind uncertainties (i.e. the same application as in Béal et al., 2010, in Sect. 4), and the ecosystem response to parameters uncertainties (Sect. 5). On the other hand, we also show examples of anamorphic transformations applied to the non-stochastic ensemble description of forecast uncertainties in current pre-operational developments for the sea ice component (Mercator system, Sect. 6) and for the ecosystem component (MyOcean project, Sect. 7). In addition, before going to the applications, the paper includes a brief summary of the algorithm (presented in a more deductive way than in Béal et al., 2010), with a quantitative discussion of the computational complexity and accuracy of the approximation (Sect. 2).

2 Anamorphosis transformations

The basic problem of the algorithm is to look for a nonlinear change of variable transforming a random variable X with known cumulative distribution function (cdf) $F(x) = p(X \leq x)$ into a new random variable Z with the target cdf $G(z) = p(Z \leq z)$. Elementary probability calculus (e.g. Von Mises, 1964) provides a general solution for the forward and backward transformations:

$$Z = G^{-1}[F(X)] \quad \text{and} \quad X = F^{-1}[G(Z)] \tag{1}$$

providing that F and G are invertible. In particular, if $Z \sim \mathcal{U}(0,1)$ is uniformly distributed on the interval $[0,1]$, with $G(z) = z$, then $x = F^{-1}(k/q)$ is the kth q-quantile of X; and if $Z \sim \mathcal{N}(0,1)$ is normally distributed, with $G(z) = \frac{1}{2}[1 + \mathrm{erf}(z/\sqrt{2})]$, then Eq. (1) defines the forward and backward Gaussian anamorphosis transformation of the random variable X (Wackernagel, 2003, chapter 33).

However, it is important to keep in mind that transforming all variables of a random vector using Eq. (1) can only ensure that the marginal distribution of each variable becomes Gaussian. This does not imply that their joint probability distribution becomes a multivariate Gaussian distribution, which is the condition required to apply linear estimation techniques. As pointed out by Wackernagel (2003), it is thus important to check in practice that at least bivariate distributions of the transformed variables become close to bi-Gaussian, so that linear inference may be close to optimal. It is the purpose of the present paper to check this in various oceanic applications, by studying how the transformation in Eq. (1), applied separately for every random variable, at every spatial location, modifies the spatial correlation structure. But before going to the applications, this section is dedicated to describing the specific algorithm that we have implemented to approximate Eq. (1) using a limited-size sample of the random variables.

2.1 Efficient approximate algorithm

In the Monte Carlo estimation methods (like the ensemble Kalman filter), the prior probability distribution for the control variables is only approximately described by a finite-size sample. The anamorphosis transformation in Eq. (1) for each control variable can thus only be approximately computed from the available sample using a nonparametric estimate $\tilde{F}(x)$ of the exact marginal cdf $F(x)$. The most simple nonparametric estimate of a probability density function (pdf) $\tilde{f}(x) = d\tilde{F}(x)/dx$ is the histogram (Izenman, 2008, chapter 4): a piecewise constant pdf $\tilde{f}(x)$, or a piecewise linear cdf $\tilde{F}(x)$. As a simple choice for the classes of the

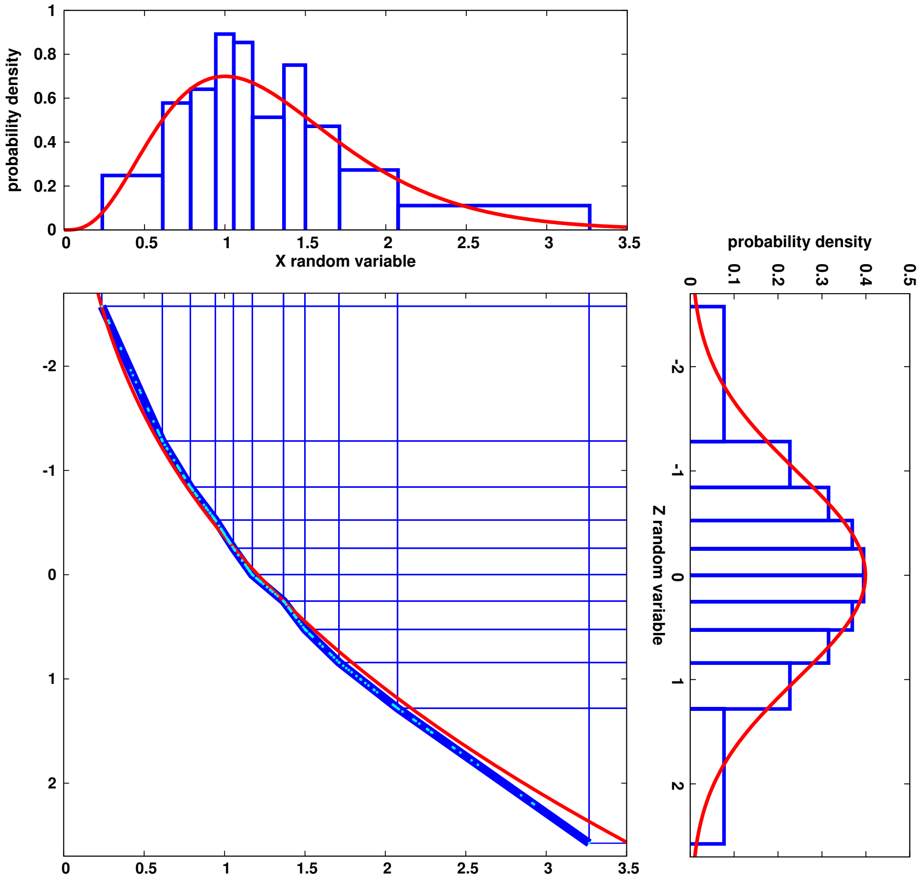

Fig. 1. Approximate piecewise linear Gaussian anamorphosis transformation (thick blue curve), remapping the deciles $\tilde{x}_k$ of a 200-member random sample of the Gamma distribution $\Gamma(k,\theta)$ (top histogram) on the Gaussian deciles z_k (right histogram), as compared to the exact transformation (in red) transforming the exact $\Gamma(k,\theta)$ (red curve superposed to the top histogram) into $\mathcal{N}(0,1)$ (red curve superposed to the left histogram).

histogram, we may use prescribed quantiles $\tilde{x}_k$, $k=1,\ldots,q$ of the input sample, i.e. such that $\tilde{F}(\tilde{x}_k)=r_k$, for a given set of r_k $(0\le r_k\le 1,\ r_k<r_{k+1})$. In this way, we can control explicitly the fraction of ensemble members $(r_{k+1}-r_k)$ in each class of the histogram.

Then, with the same level of approximation, we can use the same histogram representation of the Gaussian distribution, i.e. a piecewise linear $\tilde{G}(z)$ interpolating the true Gaussian cdf between $G(z_k)=r_k$, $k=1,\ldots,q$, so that the anamorphosis transformation in Eq. (1) is also piecewise linear:

$$\varphi_{\mathrm{forw}}(x)=\tilde{G}^{-1}\left[\tilde{F}(x)\right]=z_k+\frac{z_{k+1}-z_k}{\tilde{x}_{k+1}-\tilde{x}_k}(x-\tilde{x}_k)\quad \text{for}\quad x\in[\tilde{x}_k,\tilde{x}_{k+1}] \tag{2}$$

$$\varphi_{\mathrm{back}}(z)=\tilde{F}^{-1}\left[\tilde{G}(z)\right]=\tilde{x}_k+\frac{\tilde{x}_{k+1}-\tilde{x}_k}{z_{k+1}-z_k}(z-z_k)\quad \text{for}\quad z\in[\tilde{z}_k,\tilde{z}_{k+1}] \tag{3}$$

This approximate transformation (heuristically proposed by Béal et al., 2010) remaps the quantiles $\tilde{x}_k$, $k=1,\ldots,q$ of the input sample on the corresponding Gaussian quantiles z_k, $k=1,\ldots,q$, and interpolates linearly between them. It is bijective between the interval $[\tilde{x}_1,\tilde{x}_q]$ and $[z_1,z_q]$, providing that the quantiles $\tilde{x}_k$ are distinct: $\tilde{x}_k\neq\tilde{x}_{k+1}\ \forall k$ (see Sect. 2.3 for a discussion of the degenerate cases $\tilde{x}_k=\tilde{x}_{k+1}$, and for possible parameterizations of the tails of the probability distributions: $x\notin[\tilde{x}_1,\tilde{x}_q]$).

Example: Figure 1 shows for instance the approximate Gaussian anamorphosis transformation that is obtained with Eq. (2) using a 200-member random sample of the Gamma distribution $X\sim\Gamma(k,\theta)$, with $k=4.236$ and $\theta=0.309$ (chosen so that the mode is equal to 1, and the 95 % percentile is equal to 2.5). The classes of the histogram for X are defined using the 10-quantiles (or deciles) of the random sample: $r_k=k/q$, with $q=10$. They are remapped on the Gaussian deciles z_k (histogram on the right) using the piecewise linear transformation (blue curve), which is here not far from the exact transformation (red curve), given by Eq. (1). With this definition of r_k, there is the same number of random draws in each class of the histogram.

Computational complexity: The first reason why such a simple approximation of the Gaussian anamorphosis may be useful in practical ocean applications is that it can be performed at a numerical cost that is usually much smaller than the numerical cost of a Gaussian observational update (e.g. the analysis step of the ensemble Kalman filter). In the identification of the approximate transformation in Eq. (2), the main cost is associated to the computation of the quantiles $\tilde{x}_k$ of the input sample. If m is the size of the sample, this cost is proportional to $m \log m$, to sort the sample values. Then, if n is the size of the control vector (i.e. the number of random variables to transform), the total computational complexity to identify the functions φ_{forw} and φ_{back} in Eqs. (2) and (3) is:

$$C_{\mathrm{quantiles}} \sim nm \log m \tag{4}$$

In addition, in order to perform the observational update, one must apply the transformation to the ensemble forecast and to the observations. Each transformation requires localizing the input value among the quantiles $\tilde{x}_k$ (with complexity proportional to $\log_2 q$ with a bissection method), and then applying the corresponding linear transformation in Eq. (2) (i.e. about 3 operations). To transform the ensemble of m control vectors, together with the p observations values, and then the updated ensemble back in the original control space, this corresponds to a computational complexity of:

$$C_{\mathrm{anamorphosis}} \sim (2mn + p)(3 + \alpha \log_2 q) \tag{5}$$

where α stands for the relative numerical cost between numerical comparisons (needed to localize values in the list of quantiles) and algebraic operations (needed to compute the linear transformations). Transforming the observations simply requires applying the observation operator to the quantiles of the control vector, but if some observations are nonlinearly linked to the control vector, it may be better to augment the control vector with these observations (thus producing a problem with larger n) and transform them using their own anamorphosis transformation.

On the other hand, this simple algorithm does not require a lot of memory or disk space to store the approximate functions φ_{forw} and φ_{back}: only the quantiles of the input ensemble $\tilde{x}_k$, $k = 1, \ldots, q$ need to be stored, for a total storage of qn real values (i.e. less than the storage of the forecast ensemble itself, which requires storing mn real values). See the appendix for more details about the practical implementation of the algorithm.

2.2 Accuracy of the approximation

The second reason why such a simple approximation may be useful in practical ocean applications is that the accuracy of the approximation is generally sufficient to substantially improve the description of the marginal distributions. The accuracy of the approximation given by Eq. (2) mainly depends on the accuracy of the histogram description of $f(x)$, which is related to the size of the sample and to the definition of the classes of the histogram by the quantiles $\tilde{x}_k$. With too many quantiles, we are likely to introduce spurious features in the transformed pdf (not resolved by the available ensemble), and with too few quantiles, we will smooth out significant features. Thus, for a given distribution and a given sample size, there exists an optimal resolution of the quantiles giving the best approximation for the transformation.

For the example of Fig. 1, we computed the approximate anamorphosis transformation from the same 200-member sample and for several resolution of the histogram ($q = 3$ to 50 with regular quantile discretization: $r_k = k/q$, $k = 0, \ldots, q$). Then, we transformed the exact prior distribution $\Gamma(k, \theta)$ using these various approximations and computed the relative entropy (as a measure of the discrepancy between two pdfs, see for instance Bocquet et al., 2010) between the resulting transformed pdfs and the target transformed pdf $\mathcal{N}(0,1)$. Figure 2 (left panel) shows that there is indeed an optimal number of quantiles ($q = 9$), which is close to the choice that we made in Fig. 1 ($q = 10$). (Oscillations occur for large q because the number of ensemble members in each class of the histogram becomes too small to produce an accurate estimation of the transformation.)

Gaussian mixture: Other estimates of the transformation function can be obtained using more sophisticated nonparametric estimates of $f(x)$, for instance by approximating the unknown pdf by a mixture of Gaussian kernels (Izenman, 2008) rather than a mixture of uniform kernels (as in the histogram approximation). A common algorithm to estimate the Gaussian mixture from the available sample can be derived from the nearest neighbour method (e.g. Silverman, 1986; Izenman, 2008): each member of the sample is used as the mean of one of the superposed Gaussian pdfs, with a variance equal to the variance of the q nearest neighbours. As in the histogram approximation, there is an optimal q below which spurious features are introduced in the pdf estimate, and above which significant features are smoothed out.

Figure 2 (middle panel) shows however that this optimal q (minimizing the relative entropy) produces an estimate of $f(x)$ that is not better than the best histogram (even if the behaviour as a function of q is more regular). Moreover, the numerical cost of the transformation, requiring numerical root-finding in the integral of the superposed Gaussian pdfs [to solve the equation $\tilde{F}(x) = \tilde{G}(z)$], would be much too high to be affordable in large size applications.

Polynomial development: Another way of constructing a direct approximation of the anamorphosis transformation (described in Wackernagel, 2003) is (i) to approximate $F(x)$ by the cumulative histogram $\tilde{F}(x) = \alpha/m$ for

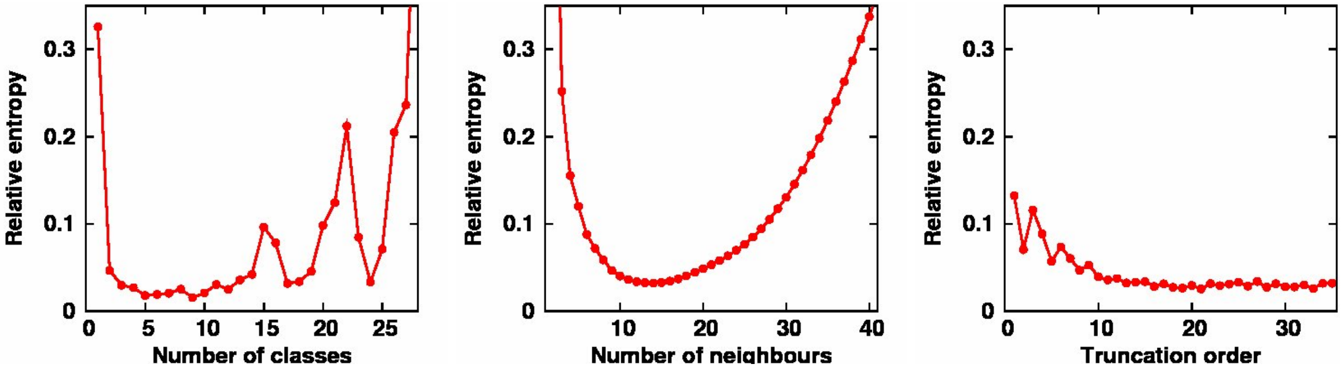

Fig. 2. Relative entropy between the transformation of the exact $\Gamma(k,\theta)$ and $\mathcal{N}(0,1)$, using various approximations of the transformation function: the histogram approximation (left panel), as a function of the number q of classes in the histogram, the Gaussian mixture approximation (middle panel), as a function of the number q of nearest neighbours, the Hermite polynomial development (right panel), as a function of the number q of superposed polynomials. In all 3 cases, the relative entropy is computed by numerical integration over the same interval $|z| < 2.576$, except in the 3rd case (polynomial development) for which the subintervals with zero density (due to the non-bijectivity of the approximation) have been removed.

$x \in [x_{(\alpha)}, x_{(\alpha+1)}]$ where $x_{(\alpha)}$, $\alpha = 1,\ldots,m$ is the ordered sample (i.e. a step function instead of a piecewise linear function in the approximation above), (ii) to deduce the corresponding transformation as $G^{-1}[\tilde{F}(x)] = G^{-1}(\alpha/m)$ for $x \in [x_{(\alpha)}, x_{(\alpha+1)}]$, or reciprocally, to construct an empirical anamorphosis transformation as $\tilde{F}^{-1}[G(z)] = x_{(\alpha)}$ for $z \in [G^{-1}(\frac{\alpha-1}{m}), G^{-1}(\frac{\alpha}{m})]$ (i.e. again a step function, which is not bijective by construction), and (iii) to interpolate this empirical anamorphosis transformation by a limited development in Hermite polynomials (see Wackernagel, 2003, for more detail about this algorithm). The q-th order Hermite development can be shown to be the best q-th order approximation (of the transformation function) in the least square sense (Wackernagel, 2003), but nothing guarantees that the polynomial interpolation will produce a bijective transformation, as it should be, so that ad hoc corrections must be supplied if problems occurs. (To avoid this problem, Simon and Bertino, 2009, linearly interpolate the step function instead of the development in Hermite polynomials.)

Figure 2 (right panel) shows the relative entropy between the transformed pdf obtained with this method and the exact pdf, as a function of the truncation order q in the development of Hermite polynomials. Again, there exists a best truncation order $q = 21$, which is not more accurate than the histogram best estimate (shown on the left panel). These results suggest that, with a moderate size sample (200 members in this example), it is not easy to do better than the simple histogram approximation, and that more sophisticated (and more expensive) algorithms, like the Gaussian mixture or the polynomial development, need a substantial increase in the sample size before producing a significant benefit.

2.3 Extensions of the algorithm

The algorithm described above is sufficient and well-conditioned as soon as (i) the cdf $F(x)$ of every control variable is invertible (so that the quantiles of the ensemble are distinct), (ii) the range of possible value for every control variable is finite (between x_1 and x_p), and (iii) the size m of the ensemble is large enough to provide a reasonable approximation $\tilde{F}(x)$ of the marginal distributions. The purpose of this section is to examine what may be done if these 3 conditions are not verified.

Probability concentrations: A cdf $F(x)$ is not invertible if it makes a vertical step at some value $x = x_c$, i.e. if there is a probability concentration for $x = x_\mathrm{c}$, with finite probability: $p(x_\mathrm{c}) = F(x_\mathrm{c}^+) - F(x_\mathrm{c}^-)$. In this case, several ensemble members may be equal to x_c [$mp(x_\mathrm{c})$ members in average] so that a subset of the quantiles (between $\tilde{x}_l$ and $\tilde{x}_\mathrm{u}$) may also be equal to x_c, and the piecewise linear approximation of the anamorphosis transformation is no more bijective (zero denominator in Eq. 2). This occurs very often in practice, especially if there is a physical constraint on the value of the random variable, so that probability may concentrate on the constraint: sea temperature equal to freezing point, zero tracer concentration (see examples in Sects. 4, 5 and 7), ice fraction equal to 0 or 1 (see example in Sect. 6), ice velocity equal to 0 (no motion),...

The most direct solution to this problem (applied in all applications below, except in the Mercator applications in Sect. 6) is to transform x_c to the middle of the step of the piecewise linear function: $\tilde{G}^{-1}[\tilde{F}(x_\mathrm{c})] = \frac{1}{2}(\tilde{x}_l + \tilde{x}_u)$. A difficulty with this simple scheme is that it can introduce spurious discontinuities in the transformed vector (for instance in the transformed ice concentrations, at the border of the ice pack in the example of Sect. 6), and it may be preferable to restore the bijectivity of the transformation by introducing an artificial slope in the function. A simple way to do it is to replace the quantiles $\tilde{x}_l$ to $\tilde{x}_u$ (all equal to x_c) by interpolating them between $\tilde{x}_{l-1}$ and $\tilde{x}_{u+1}$ (between $\tilde{x}_1$ and $\tilde{x}_{u+1}$ if $l=1$, or between $\tilde{x}_{l-1}$ and $\tilde{x}_q$ if $u=q$). This can improve the continuity and the quality of the linear estimates in the transformed space (see Sect. 6), at the price of a slight spreading of the backward transform around the concentration value x_c (above x_c if $l=1$, or below x_c if $u=q$).

Tails of the distribution: Since the range of possible values for the Gaussian random variable Z is between $-\infty$ and $+\infty$, the backward transformation in Eq. (3) must also specify how to transform $z<z_1$ and $z>z_q$. If the range of possible values for the original random variable X is finite between $x_{\min}$ and $x_{\max}$, and fully resolved by the available ensemble (so that $\tilde{x}_1 = x_{\min}$ and $\tilde{x}_q = x_{\max}$), then we can be certain that the cumulated probability corresponding to $z<z_1$ and $z>z_q$ is concentrated at $x=x_{\min}$ and $x=x_{\max}$, so that the backward transformation may be written:

$$\varphi_{\mathrm{back}}(z)=\tilde{x}_1 \quad \text{for} \quad z<z_1 \tag{6}$$

$$\varphi_{\mathrm{back}}(z)=\tilde{x}_q \quad \text{for} \quad z>z_q \tag{7}$$

But if the range between $x_{\min}$ and $x_{\max}$ (possibly infinite) is not fully resolved by the available ensemble, a solution must be provided to map $[-\infty, z_1]$ on $[x_{\min}, \tilde{x}_1]$, and $[z_q, \infty]$ on $[\tilde{x}_q, x_{\max}]$.

The most simple parameterization of the tails of $F(x)$ (used in all applications below) is to assume zero probability outside the range of the ensemble forecast (as in Béal et al., 2010). Again, this corresponds to assuming that the cumulated probability corresponding to $z<z_1$ and $z>z_q$ is concentrated at $x=x_{\min}$ and $x=x_{\max}$, so that the backward transformation is approximated by Eqs. (6) and (7). On the other hand, any x found outside of the interval $[\tilde{x}_1, \tilde{x}_q]$ is viewed as impossible and transformed as the closest value:

$$\varphi_{\mathrm{forw}}(x)=z_1 \quad \text{for} \quad x<\tilde{x}_1 \tag{8}$$

$$\varphi_{\mathrm{forw}}(x)=z_q \quad \text{for} \quad x>\tilde{x}_q \tag{9}$$

Parameterizing the tails of $F(x)$ by probability concentrations at $\tilde{x}_1$ and $\tilde{x}_q$ means that the resulting transformation cannot be bijective outside of the interval $[\tilde{x}_1, \tilde{x}_q]$. However, if the available ensemble is large enough and consistently sampled (without bias) from the prior probability distribution, these tails must correspond to a very small cumulated probability. Moreover, if little is known about the extreme behaviour of the system, Eqs. (6) to (9) may be a safe way of avoiding any kind of extrapolation outside the range of values that has been explored by the ensemble.

More sophisticated assumptions about the tails of $F(x)$ can nevertheless be easily implemented. See for instance Simon and Bertino (2009) for a Gaussian parameterization (requiring that $x_{\min}$ or $x_{\max}$ be infinite).

Sample enrichment: In many practical applications, it may be very expensive to increase the ensemble size m until the accuracy of the approximation is sufficient to improve (or at least not deteriorate) the Gaussianity of the marginal distributions. In such circumstances, and providing that $F(x)$ is slowly varying in space, a better accuracy of $\tilde{F}(x)$ at a given location $\mathbf{x}$ can certainly be obtained (for a moderate size m) by augmenting the sample that is available at $\mathbf{x}$, with the samples that are available in the neighbourhood of $\mathbf{x}$ (possibly with a decreasing weight as a function of the distance from $\mathbf{x}$). However, the definition of this neighbourhood (which should decrease with m) introduces a subjective parameter in the algorithm, which can only be optimized by checking the accuracy of the results. This is why no enrichment of the sample is used in the applications below (except in the Mercator application in Sect. 6), where we preferred to stick to the theoretical formulation (converging for $m \to \infty$) of separate transformations for distinct random variables (Wackernagel, 2003).

Finally, it is important to remark that such a spatial extension of the sample is by no way necessary to ensure the spatial smoothness of the approximate solution described in Sect. 2.1. If all ensemble members $x_{(\alpha)}$ are spatially smooth, their quantiles $\tilde{x}_k$ and thus the anamorphosis transformation in Eqs. (2) and (3) will be spatially smooth as well (see applications below), so that no spurious discontinuity is introduced in the multivariate probability distribution. On the contrary, one should certainly be careful enough to check that the sample extension described above does not smooth out real discontinuities (or sharp gradients) from the statistics. Again, what really matters to apply linear estimation methods is that the joint probability distribution for all control variables, at every spatial location $\mathbf{x}$, is better described by a multivariate Gaussian distribution if the nonlinear change of variables proposed in Eq. (2) is applied. It is precisely the purpose of the following examples to show that such local anamorphic transformations may yield a far better model for various kind of ocean uncertainties.

2.4 Effect on correlations

However, since the examples given in the following sections are mainly dedicated to illustrate the effect of anamorphic transformations on spatial correlations, it is certainly useful to provide first a summary of the theoretical background explaining the effect that can be expected. For that purpose, we assume that we have two non-Gaussian random variables X_1 and X_2 (with marginal cdfs F_1 and F_2) that have been transformed into the Gaussian variables Z_1 and Z_2 (with the same cdf G). First of all, it is important to remember that, since the transformations are invertible, there is no loss of information induced by the anamorphosis, and the statistical dependence (in a general sense) between the random variables remains unchanged, i.e. the reduction of entropy gained from the knowledge of the other variable (i.e. the mutual information I) remains the same:

$$\begin{aligned} I(X_1,X_2) &= H(X_2)-H(X_2|X_1)=H(Z_2)-H(Z_2|Z_1) \\ &= I(Z_1,Z_2) \end{aligned} \tag{10}$$

which can easily be verified by introducing the change of variables in the definition of entropy $[H(X_2)]$ and conditional entropy $[H(X_2|X_1)]$. Consequently, it is only the effect of anamorphic transformations on *linear* correlations that we are going to investigate, since this is the only kind of correlation that can be described by a Gaussian model.

A first insight into this problem can easily be obtained by remarking that, if there exists separate bijective transformations for X_1 and X_2 transforming their joint non-Gaussian distribution into a bi-Gaussian distribution for Z_1 and Z_2, then the anamorphic transformation given by Eq. (1) provides the required transformations. This is obvious since the marginal pdfs of a bi-Gaussian distribution are both Gaussian, and the only backward anamorphosis (except for any unimportant additional linear change of variable) transforming the Gaussian marginal pdf for Z_1 and Z_2 into the right marginal pdfs for X_1 and X_2 is the one given by Eq. (1). In this ideal case, the mutual information is related to the linear correlation coefficient $\rho_{Z_1Z_2}$ between the transformed variables (e.g. Cover and Thomas, 2006, chapter 8) by:

$$I(X_1,X_2)=I(Z_1,Z_2)=-\frac{1}{2}\ln(1-\rho^2_{Z_1Z_2}) \tag{11}$$

A particular case of this ideal situation occurs if the variable X_1 and X_2 are perfectly correlated along a monotonic nonlinear curve (i.e. the ideal situation to estimate X_2 from an observation of X_1, but in which linear estimation methods can be very inaccurate). In this case, by transforming the two marginal pdfs into Gaussian pdfs, the anamorphic transformations also transform the nonlinear curve into a straight line (so that the two marginal pdfs can be simultaneously Gaussian). The nonlinear dependence between X_1 and X_2 (resulting from their non-Gaussian behaviour) is fully transformed into a linear dependence, which is then perfectly described by the bi-Gaussian pdf (i.e. linear estimation methods become truly optimal). In this particular case, the linear correlation coefficient, which only imperfectly described the perfect nonlinear dependence between X_1 and X_2, is always amplified by the transformation ($|\rho_{X_1X_2}|<|\rho_{Z_1Z_2}|\simeq 1$), as a direct consequence of the transformation of the nonlinear curve into a straight line. This first explanation thus covers all situations in which $|\rho_{Z_1Z_2}|$ is close to 1, because this means that all transformed values are aligned close to a straight line (as a result of the transformation of a nonlinear regression curve into a straight line). This kind of behaviour is what is observed for spatial correlations in most examples described in Sects. 3 to 7.

Nevertheless, it is important to stay aware that, in general, only the marginal distributions $p(Z_1)$ and $p(Z_2)$ are ensured to be Gaussian, and that assuming that $p(Z_1,Z_2)$ is bi-Gaussian is only an approximation. This is why, in this case, it is much more difficult to make general mathematical statements about the transformation of linear correlations. A useful way to understand how linear correlations are modified by the transformation $X_1,X_2\rightarrow Z_1,Z_2$ is to observe that the linear coefficient between the transformed variables Z_1 and Z_2 corresponds to a nonparametric measure of correlation between the original variables X_1 and X_2, because there is an abundant statistical literature explaining the advantages of nonparametric correlations as compared to linear correlations (Hollander and Wolfe, 1973; Corder and Foreman, 2009). In summary, the two main advantages are (a) that they are more adequate to see a nonlinear dependence between random variables (for the same kind of reason as in the ideal case described above), and (b) that they are more robust to the presence of outliers in the data. These two cases correspond to the situations in which the linear correlation can provide an inaccurate representation of the dependence between the random variables (as illustrated in the examples of Anscombe, 1973). And the basic reason underlying these two improvements is the derivation of variables that are identically distributed (Z_1 and Z_2 are both normal in our case).

The oldest and most simple example of a nonparametric measure of correlation is the rank correlation (Spearman, 1904; Kendall, 1962), which is defined as the linear correlation between the rank of each member in the ensemble. Hence, this corresponds to computing a linear correlation between uniform sets of integers between 1 and m, which is thus close to computing a linear correlation after a uniform anamorphosis (i.e. with a uniform target pdf), instead of a Gaussian anamorphosis. (This is only approximate because, unlike uniform anamorphosis, the computation of the rank is not invertible, so that there is a small loss of information in the operation.) The close similarity between the rank correlation between X_1 and X_2 and the linear correlation between Z_1 and Z_2 was already discussed in Béal et al. (2010), and it is further illustrated here in the example of Sect. 4 (Fig. 6). This property that the linear correlation coefficient $\rho_{Z_1Z_2}$ between the transformed variables corresponds to a nonparametric measure of correlation between the original variables (similar to the rank correlation) is the fundamental reason explaining the improvement of the correlation structure that is described in the rest of the paper. By this, we will always mean that the resulting nonparametric measure of correlation is more adequate to see a nonlinear dependence between the random variables and more robust to the presence of outliers in the data (as already observed in other applications of anamorphosis in Geostatistics, see Chilès and Delfiner, 1999).

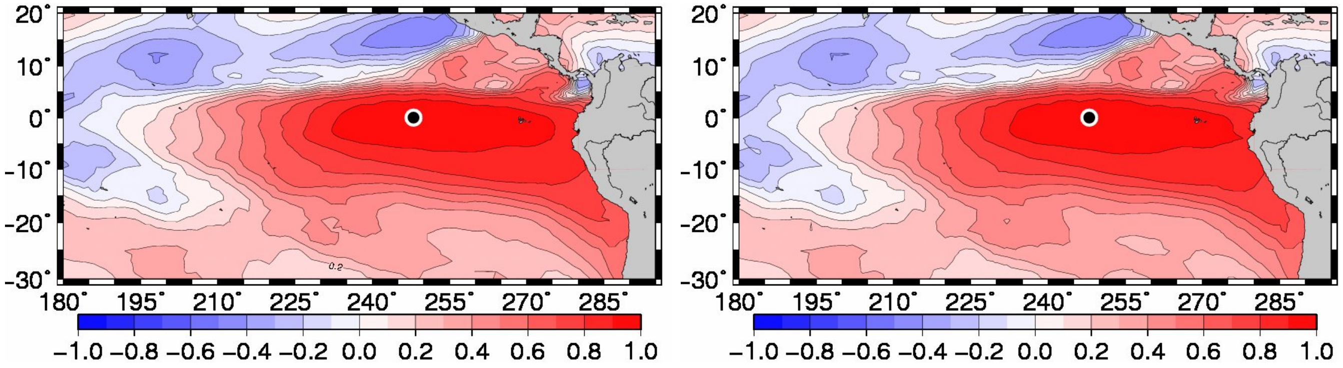

Fig. 3. SST horizontal correlation structure with respect to SST at 114° W 0° N (Eastern Equatorial Pacific), without anamorphosis (left panel), and after local anamorphosis transformations (right panel).

3 Mixed layer response to atmospheric forcing uncertainties

As a first example, we study the stochastic response of the ocean mixed layer to uncertainties in the atmospheric parameters that are used to define the surface boundary condition of the ocean model (i.e. the momentum, heat and fresh water fluxes). In many respects, the ensemble model forecast that we use here to illustrate the effect of anamorphosis transformations is similar to the ensembles that are used in Skandrani et al. (2009) to estimate corrections in the atmospheric parameters using oceanic observations (without anamorphosis), because (i) we use the same low resolution global ocean configuration (ORCA2) of the NEMO-OPA model (Madec and Imbard, 1996), with a 2° × 2° ORCA type horizontal grid and 31 z-coordinate levels along the vertical, and (ii) the random parameters perturbations are drawn from a Gaussian probability distribution with zero mean and a covariance derived from their natural variability. However, the ensemble that we describe here (performed by Meinvielle, 2011) is also somewhat different because (i) the reference atmospheric parameters are obtained from the ERA-interim dataset instead of NCEP, with the objective (not discussed here) of estimating parameter corrections for long term model simulations (The DRAKKAR Group, 2007), (ii) the parameter perturbations now include the wind, and are assumed constant over monthly periods, rather than weekly periods (to estimate lower frequency parameter corrections), and (iii) the covariance of the perturbations is set to the covariance of the ERA-interim monthly means (from 1989 to 2007) for the 3 months surrounding the month of interest, rather than the full covariance of the parameter variability in Skandrani et al. (2009). In the following, we focus our study to the one-month and 200-member ensemble model forecast that is produced for January 2004, and we look at the mixed layer response, averaged over the one-month time period, in terms of sea surface temperature (SST), sea surface salinity (SSS) and mixed layer depth (MLD).

Figure 3 shows for instance the resulting ensemble correlation structure with respect to SST at 114° W 0° N (Eastern Equatorial Pacific), without anamorphosis (left panels), and after local anamorphosis transformations (right panels) based on the deciles of the ensemble forecast (as in Fig. 1). What we observe is that the SST horizontal correlation structure is (almost) not modified by the local transformations. This occurs here because the ensemble model response to Gaussian parameter perturbations is already very close to Gaussian, so that the ensemble deciles for SST, at every location, are all remapped on the deciles of $\mathcal{N}(0,1)$ along a straight line. Conversely, this means that the approximate algorithm described in Sect. 2, with piecewise linear transformations based on a histogram description of the probability distributions, is accurate enough (with 200 members) to faithfully preserve the linear correlation structure between random variables that are already close to Gaussian. This Gaussian behaviour is also the reason why Skandrani et al. (2009) were able to infer relevant parameter corrections from SST (and SSS) using a Gaussian observational update algorithm (complemented by the truncated Gaussian assumption of Lauvernet et al., 2009, to avoid extreme and nonphysical corrections).

However, the situation becomes different if we look at the ensemble model response in terms of MLD. Figure 4 shows for instance the correlation structure with respect to MLD at the same location (114° W 0° N in Eastern Equatorial Pacific), without anamorphosis (left panel), and with the same local anamorphic transformation as above (right panels). What we observe is that, for both displayed variables (MLD and SST), the horizontal correlation patterns are not really altered by the local transformations (same smoothness, same shape, same kind of anisotropy), but the correlation radius is substantially increased in all directions: the area inside which the correlation (or anticorrelation) is above 70 % is increased by 50 % for MLD and 62 % for SST. This means that the MLD response to Gaussian parameter perturbations is not Gaussian, as illustrated in Fig. 5 (left panel) by a

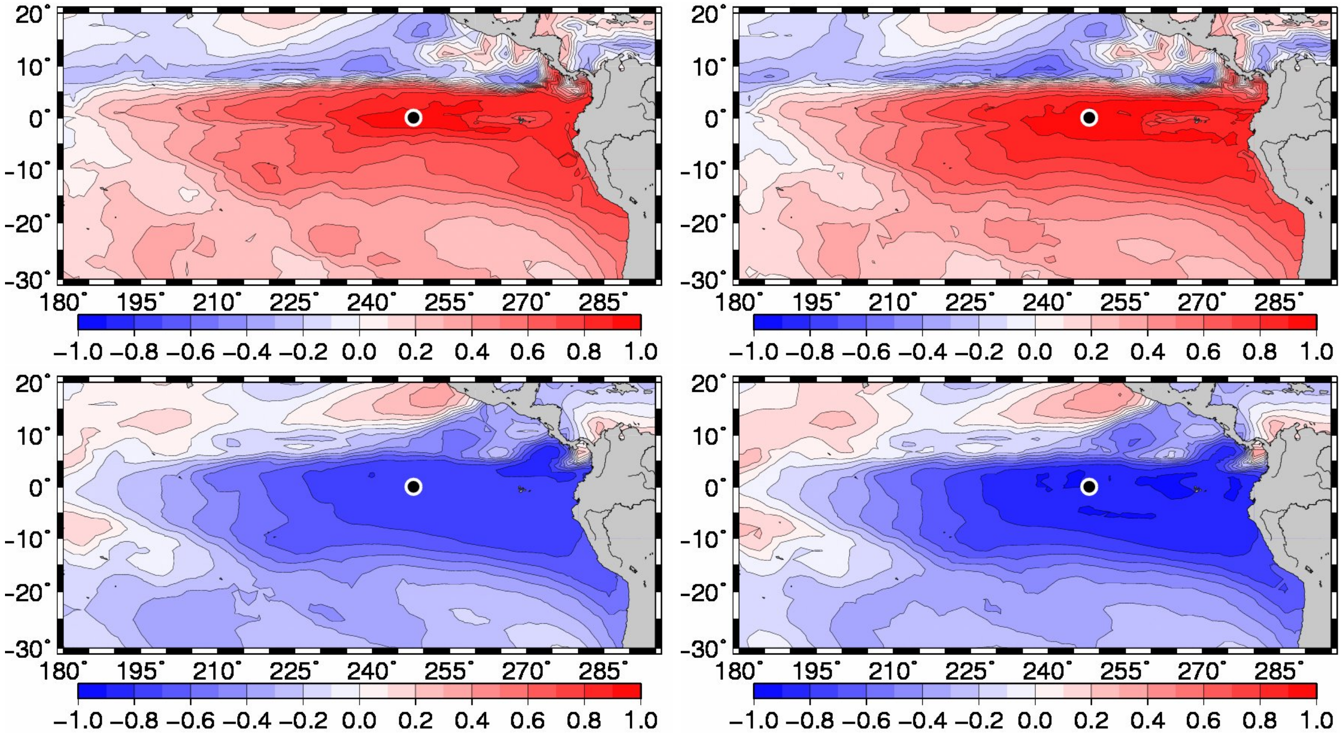

Fig. 4. MLD (top panels) and SST (bottom panels) horizontal correlation structure with respect to MLD at 114° W 0° N (Eastern Equatorial Pacific), without anamorphosis (left panels), and after local anamorphosis transformations (right panels).

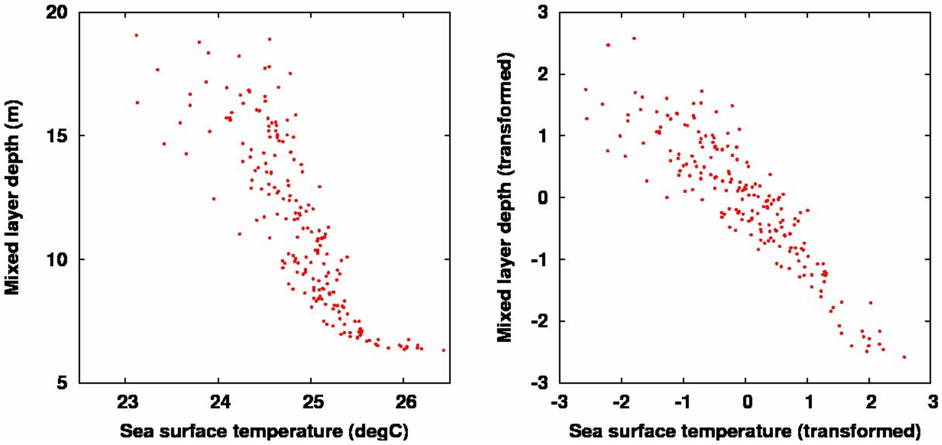

Fig. 5. Scatterplot of MLD vs. SST at 114° W 0° N (Eastern Equatorial Pacific), without anamorphosis (left panel), and after local anamorphosis transformations (right panel).

scatterplot of MLD vs. SST at 114° W 0° N. As a consequence, the joint distribution of MLD and SST cannot be bi-Gaussian, as visually obvious from the clear nonlinearity of the regression line (i.e. the line of maximum MLD probability for every given SST). In the transformed variables (Fig. 5, right panel), even if the marginal distribution for each variable is now close to Gaussian (by construction), the joint distribution is still not bi-Gaussian (larger MLD dispersion for small SST than for large SST). But at least the regression line is now close to linear, with the direct consequence of increasing the linear correlation coefficient. This phenomenon explains why the spatial correlation structure can only be improved by consistent local anamorphic transformations, even if the algorithm is not perfectly accurate (as the piecewise linear approximation). The improvement of the MLD spatial correlation structure also suggests that anamorphosis transformations might be an interesting ingredient to obtain better MLD climatologies, enhancing the accuracy of the linear

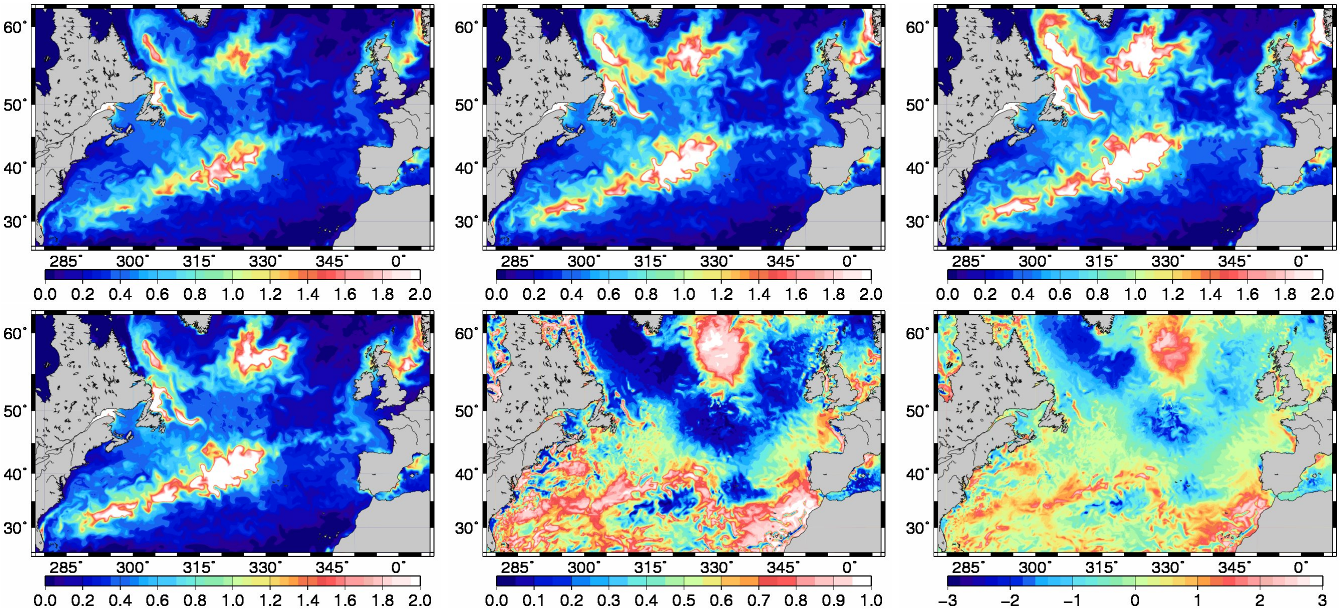

Fig. 6. Deciles of the ensemble for phytoplankton (top panels) corresponding (from left ot right) to $r_k = 0.2$, 0.5 (median) and 0.8, and illustration of one of the ensemble members (bottom panels): the phytoplankton map (left panel), the rank in the ensemble (middle panel), and the transformed map (right panel) after Gaussian anamorphosis.

estimation methods and the description of the final product (by the median and a set of quantiles, rather than the usual minimum variance estimate, which is not really meaningful in this case).

This first example already illustrates the two main conclusions of this paper about the effect of local anamorphic transformations on the spatial correlation structure: (i) the transformation is accurate enough to faithfully preserve the correlation structure if the joint distribution is already close to Gaussian, and (ii) the transformation has the general tendency of increasing the correlation radius as soon as the spatial dependence between random variables becomes nonlinear. With the next examples, we further investigate the same effects in presence of the more complex and heterogeneous non-Gaussian behaviours that may occur in ecosystem or sea-ice models.

4 Ecosystem response to wind uncertainties

As a second example, we study the stochastic response of a coupled physical-biogeochemical model (CPBM) of the North Atlantic to uncertainties in the wind forcing. For that purpose, we use the same 200-member ensemble forecast as in Béal et al. (2010): (i) the CPBM (originally developed by Ourmières et al., 2009) couples a 1/4° resolution circulation model of the North Atlantic (a Drakkar configuration of the NEMO/OPA model, The DRAKKAR Group, 2007) with the LOBSTER (LOcean Biogeochemical Simulation Tools for Ecosystem and Resources, Lévy et al., 2005) biogeochemical model, with 6 prognostic variables in the euphotic layer: phytoplankton (PHY), zooplankton (ZOO), nitrate (NO_3), ammonium, detritus, and semi-labile dissolved organic nitrogen; (ii) the ensemble forecast is initialized at the beginning of the spring bloom on 15 April 1998, using the model simulation described in (Ourmières et al., 2009); and (iii) the random wind perturbations are sampled from a Gaussian probability distribution, with zero mean and a covariance derived from the ERA40 variability (during the 3 months centered on 15 April, with a superimposed 4-day decorrelation times scale, see Béal et al., 2010, for more details). However, whereas the study by Béal et al. (2010) was exclusively focused on the multivariate response of the coupled model at given horizontal locations (with or without anamorphosis transformations, and for several forecast timescales between 1 and 30 days), we here complement their work, by documenting the effect of the local anamorphic transformations on the horizontal correlation structure (in the 4-day forecast only).

In the ensemble forecast, the main impact of the random wind perturbations on the ecosystem results from the deepening and shallowing of the mixed layer, which modifies the nutrient supply and thus the primary production in the euphotic layer. This mechanism produces a quite heterogenous response in terms of phytoplankton concentration, as illustrated in Fig. 6 by three deciles of the ensemble (corresponding to $r_k = 0.2$, 0.5 and 0.8, top panels) and one of the ensemble members (bottom panels). The wind can indeed only trigger a large ensemble dispersion (i.e. large differences between the deciles, in the top panels) in areas where

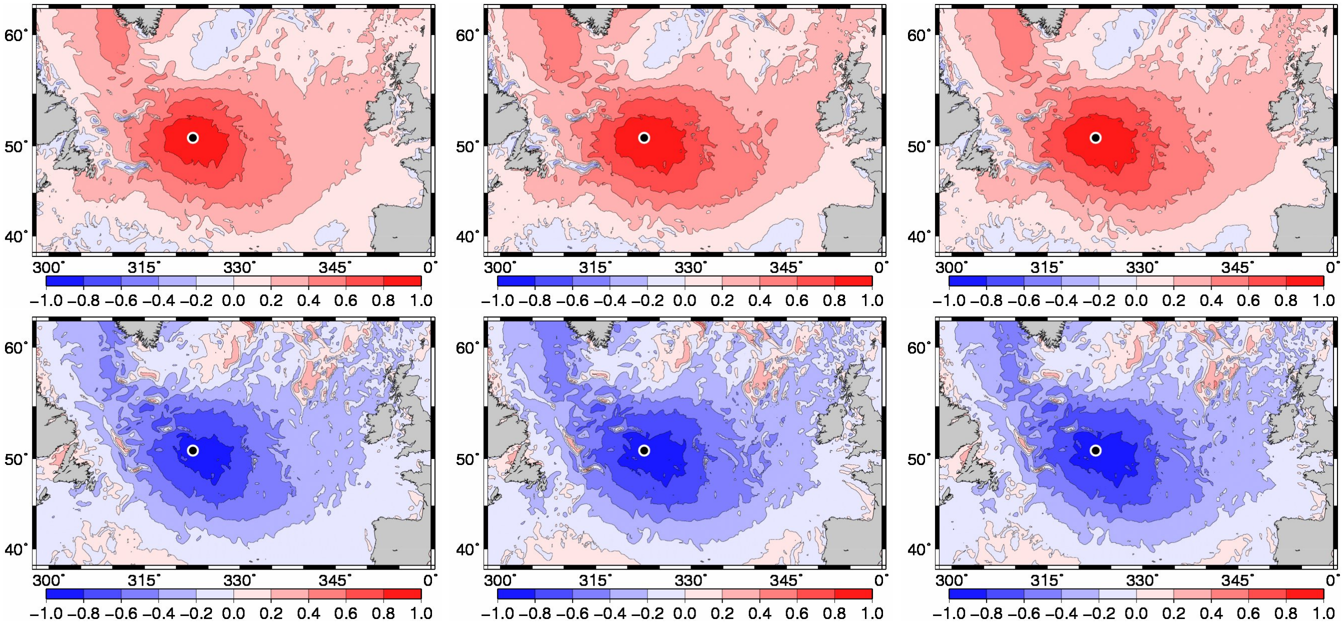

Fig. 7. Phytoplankton (top panels) and nitrate (bottom panels) horizontal correlation structure with respect to phytoplankton at 37.5° W 50.8° N (North Atlantic), as obtained for the original variables (left panels), their local rank in the ensemble (middle panels), and the transformed variables after Gaussian anamophosis (right panels).

the spring bloom has already started, like primarily the Gulf Stream pathway, the Irminger Sea and the Western half of the Labrador Sea, and secondarily, the Northern half of the North Sea, the Gulf of Lions and the Bay of Biscay. Conversely, in the areas where the primary production is weak (as in the subtropical gyre and in the Norwegian Sea), it remains weak, whatever the wind perturbations.

Furthermore, one particular ensemble member (Fig. 6, bottom left panel) may be well below the median in some regions (e.g. in the Labrador Sea) and well above the median in other regions (e.g. in the Irminger Sea). This phenomenon is more obvious if we look at the rank of this ensemble member in the ensemble forecast (Fig. 6, bottom middle panel). More precisely, what is shown is the rank divided by the ensemble size (to be between 0 and 1), which corresponds to the local anamorphic transformation of the ensemble member using the uniform distribution $\mathcal{U}(0,1)$ as a target distribution. For instance, a value below 0.2 means below the second decile ($r_2 = 0.2$), a value below 0.5 means below the median ($r_5 = 0.5$), etc. In this figure, we can see immediately where this ensemble member is high or low with respect to the others (compare the rank in the Labrador Sea and in the Irminger Sea), even in regions where the dispersion of the ensemble is very small, as along the coast of Africa or in the Southern half of the North Sea. See also how the high rank region in the Irminger Sea (i.e. with a production well above the ensemble median) embeds indifferently areas of high primary production and areas of low production, as a result of a strongly positive wind anomaly covering the whole region. The rank may thus better translate the effect of a homogeneous perturbation, which is masked in the original variable by the heterogeneity of the ecosystem dynamics. And from the local rank (Fig. 6, bottom middle panel) to the local Gaussian anamorphic transformation of the same ensemble member (Fig. 6, bottom right panel), there is nothing but a global anamorphosis transforming $\mathcal{U}(0,1)$ into $\mathcal{N}(0,1)$. The figure thus looks very similar, with the same nonlinear change of variable at every grid point (we could have kept the same figure, with a nonlinear labelling of the colorbar).

Figure 7 illustrates the effect of these transformations on the PHY (top panels) and NO_3 (bottom panels) horizontal correlation structure with respect to PHY at 37.5° W 50.8° N (North Atlantic), as obtained for the original variables (left panels), their local rank in the ensemble (middle panels), and the transformed variables after Gaussian anamophosis (right panels), based on the deciles of the ensemble forecast. The first thing that we observe is that, despite of the deep changes in the horizontal structure of each ensemble member (illustrated in Fig. 6, bottom panels), the general shape of the correlation is still not much altered by the transformations. A linear measure of correlation (Fig. 7, left panels) is already quite good in this case, because it is not influenced by the heterogeneity of the ensemble variance, which is here the main reason for the changes in the horizontal structure of the ensemble members observed in Fig. 6 (bottom panels). Going to a nonlinear measure of correlation (like the rank correlation, in the middle panels of Fig. 7) is only useful if the transformation can help linearizing the regression line

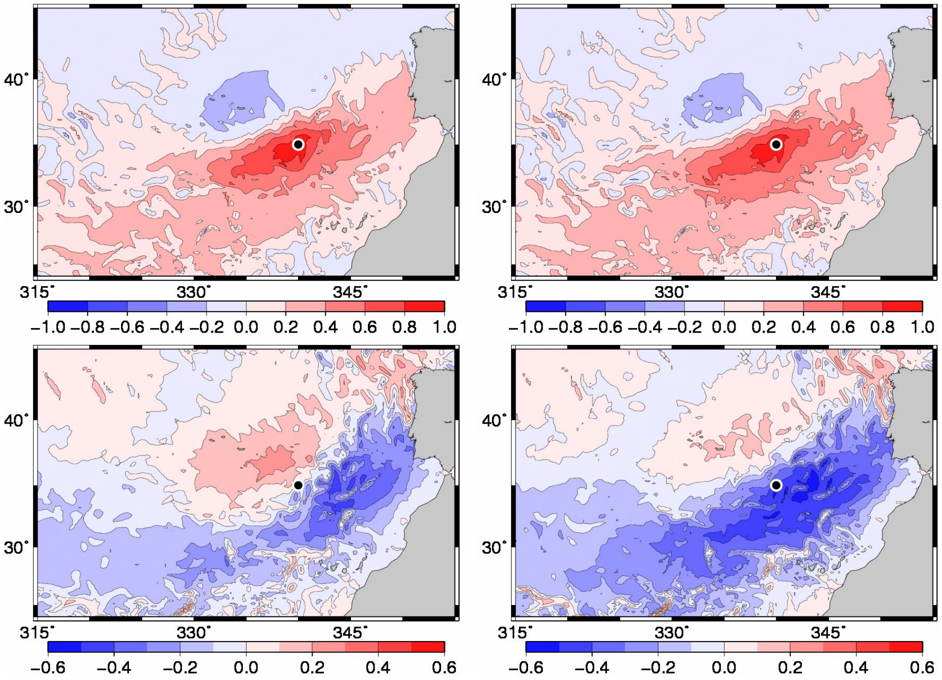

Fig. 8. Phytoplankton (top panels) and nitrate (bottom panels) horizontal correlation structure with respect to phytoplankton at 20° W 35° N (North Atlantic), without anamorphosis (left panels), and after local anamorphosis transformations (right panels).

between the two random variables (as illustrated in Fig. 5). The rank correlation was indeed introduced by Spearman (as explained by Von Mises, 1964) to produce this effect and thus to go beyond the linear correlation coefficient (of Pearson), as a measure of the (nonlinear) dependency between random variables. Furthermore, since the linear correlation structure after a local Gaussian anamorphosis is very similar to rank correlation (compare right and middle panels in Fig. 7), this explains why the correlation radius is generally increased by the transformation (compare with left panels, in which the area with a correlation above 80 % is about 26 % smaller for PHY and 35 % smaller for NO_3). The same kind of phenomenon can be observed in Fig. 8, showing the same result at 20° W 35° N, except that the rank correlation is not shown anymore since it is always very similar to the linear correlation structure after Gaussian anamorphosis. However, we can see that here, the NO_3 horizontal correlation structure (bottom panels) is deeply modified by the transformation, becoming more similar (in shape and extension, but with the opposite sign) to the PHY horizontal correlation structure (top panels). This is also related to the improvement of the correlation between NO_3 and PHY at every horizontal location (which was described in Béal et al., 2010), and further supports the idea that local anamorphic transformations may substantially increase the benefit that can be expected from ocean colour observations in the multivariate estimation of the state of the ecosystem.

5 Ecosystem response to ecosystem parameters uncertainties

As a third example, we study the stochastic response of the same CPBM to uncertainties in the parameterization of the ecosystem model. For that purpose, we use the 200-member ensemble that has been performed by Doron et al. (2011) to evaluate the potential of ensemble methods to estimate a few ecoystem parameters using ocean colour observations (with or without anamorphosis). This ensemble forecast is identical to that described in the previous section (same model, same forcing, same initial condition), except that the random perturbations are applied to a few ecosystem parameters rather than to the wind forcing. Three rate parameters are assumed uncertain in the ensemble forecast: (i) the maximum growth rate of phytoplankton, (ii) the maximum grazing rate of phytoplankton by zooplankton, and (iii) the phytoplankton mortality rate. The uncertainties for these three parameters are assumed independent and constant over each of the 13 North Atlantic biogeochemical provinces (defined by Longhurst, 1995), which makes a total of $3 \times 13 = 39$ independent random parameters. And the probability distribution for each of these parameters is assumed to be a Gamma distribution, with a mean equal to the default parameter value in the LOBSTER model, and a 95 % percentile equal to 2.5 times the mode of the distribution (as in Fig. 1, see Doron et al., 2011, for more details). In the following, we describe the correlation structure of the model response to these

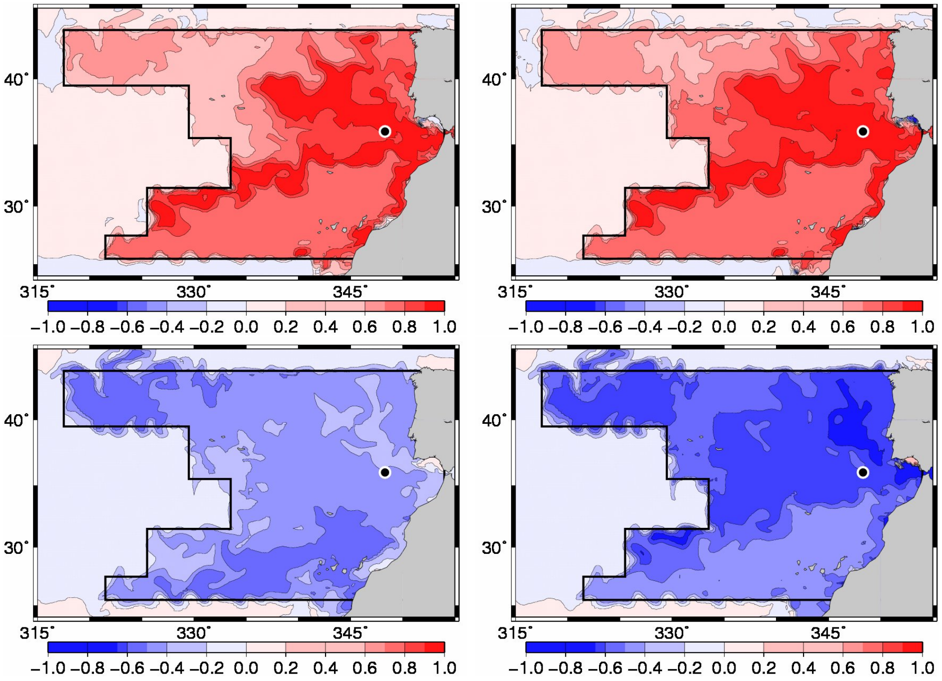

Fig. 9. Phytoplankton (top panels) and nitrate (bottom panels) horizontal correlation structure with respect to phytoplankton at 11.7° W 36° N (North Atlantic), without anamorphosis (left panels), and after local anamorphosis transformations (right panels).

uncertainties after a 1-month ensemble forecast (instead of a 4-day forecast in the previous example).

Figure 9 shows for instance the PHY (top panels) and NO_3 (bottom panels) horizontal correlation structure with respect to PHY at 11.7° W 36° N (in the Longhurst province west of Spain and North Africa), as obtained without anamorphosis (left panels) and after local anamorphosis transformations (right panels), based on the deciles of the ensemble forecast. The first thing to observe is that the correlation is mostly significant inside the Longhurst province (materialized by the black line) with constant parameters perturbations, which means (i) that the ensemble size is sufficient to decorrelate independent behaviours, and (ii) that, even after 1 month, the effect of the parameter uncertainties is here mainly local (the main exceptions being the intense mesoscale activity in the North-Western corner of the province, and the southward advection along the coast of Africa). However, inside the Longhurst province, the response of the ecosystem to the homogeneous parameters uncertainties is far from being the same everywhere, as a result of the heterogeneity of the initial condition and physical forcing. It is also clearly nonlinear, in view of the strong impact of the anamorphosis transformation on the horizontal correlation structure. As in Fig. 8, the NO_3 correlation structure becomes very similar (with an opposite sign) to the PHY correlation structure (Fig. 9, right panels), even though without anamorphosis (Fig. 9, left panels), the two variables were only weakly correlated.

Figure 10 shows the same kind of result as Fig. 9 in the Longhurst province covering the Caribbean Sea and the Gulf of Mexico, with a reference point located at 86° W 23.8° N in the inside of the Loop Current. Here, the impact of advection is more obvious: (i) along the Eastern coast of Florida, where the effect of the parameter perturbation inside the Longhurst province (delimited by the black line) is advected by the Gulf Stream, and (ii) in the Gulf of Mexico, where the ecosystem response to the parameters uncertainties decorrelates across the front defined by the Loop Current. However, even if the heterogeneity of the ecosystem behaviour across the Loop Current is clearly due to differences brought by advection, the decorrelation across the front also results from the nonlinearity of the ecosystem response to the same parameters perturbations. This is why a nonlinear measure of correlation (i.e. the linear correlation coefficient for the transformed variables, in the right panels) can be much larger than the linear correlation coefficient (for the original variables, in the left panels), going from below 0.4 to above 0.6 for PHY (the opposite sign for NO_3) in a large part of the Gulf of Mexico. It is also interesting to remark the modifications along the Western coast of the Gulf of Mexico, where a zero linear correlation transforms either to (i) a negative correlation with PHY and a positive correlation with NO_3 in the Southern half of the coastal band, (ii) a negative correlation with both PHY and NO_3 in the Northern half, or (iii) a positive correlation with PHY and a negative correlation with NO_3 (as in the rest of the domain) at the mouth of the Rio Grande. (Here, it

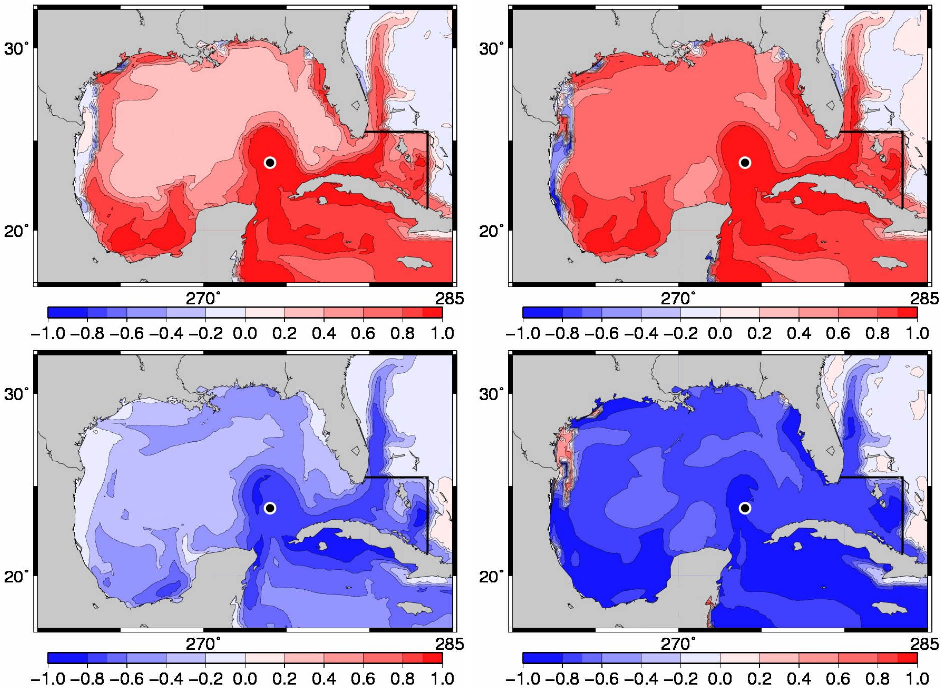

Fig. 10. Phytoplankton (top panels) and nitrate (bottom panels) horizontal correlation structure with respect to phytoplankton at 86° W 23.8° N (Gulf of Mexico), without anamorphosis (left panels), and after local anamorphosis transformations (right panels).

must be remembered that, even if these large long-range correlations are certainly meaningful, they cannot be expected to describe real model errors, because they correspond to a very simple assumption, in which parameter errors are assumed constant over the whole Gulf of Mexico.)

All these increases of linear correlation (or anticorrelation) contribute to simplify the Gaussian description of the uncertainties (in the transformed variables vs. the original variables), by concentrating a larger fraction of the total variance in a smaller dimension subspace, thus reducing the number of degrees of freedom that must be controlled to obtain a given accuracy. This simplification is one of the main reasons for which local anamorphic transformations were so helpful in the work of Doron et al. (2011) to estimate the 39 unknown parameters from ocean colour observations (in a twin experiment approach, without localization of the ensemble covariance).

6 Modelling ice forecast uncertainties

In this section, we are moving to another class of examples, in which non-stochastic ensembles are used to describe forecast uncertainties. In many situations indeed, the forward model is too expensive to allow the explicit Monte Carlo exploration of the uncertainties. Assumptions are then needed to produce the required ensemble of model states, using for instance an appropriate sample of the past system variability. The purpose of this section (and of Sect. 7) is to show that, even in such a case, local anamorphic transformations may be useful to go beyond the Gaussian model.

As a first example of this kind, we study the non-stochastic ensemble description of sea-ice forecast uncertainties that is currently tested for assimilating sea-ice observations in the Mercator/MyOcean operational system. To construct the ensemble, it is assumed that the forecast uncertainties have the same statistics as the combined effect of the forward model short term and interannual variabilities. More precisely, to describe the uncertainties at a given date (e.g. 15 June 2011), we sample a past interannual free model simulation (17 years, between 1991 and 2007) every 3 days in a running window of ±66 days around that date (thus retaining 44 model states, every year), which make an ensemble of size $m = 17 \times 44 = 748$ model states. This assumption means that we do not try to resolve anything else than the seasonal cycle in the description of the uncertainties. This might look quite crude if we forget that this is applied to a $1/4°$ resolution global configuration of the NEMO model, and already tested with a $1/12°$ resolution prototype. The size of these systems makes truly stochastic solutions (with sufficient ensemble size) unaffordable with present-day computational facilities, so that the above solution can actually be considered as quite sophisticated. In the following, we focus our study on the resulting description of the uncertainties (as obtained from the $1/4°$ resolution model) for the ice fraction f, which is the (well-observed) model variable giving the fraction of the ocean that is covered by sea-ice. It is

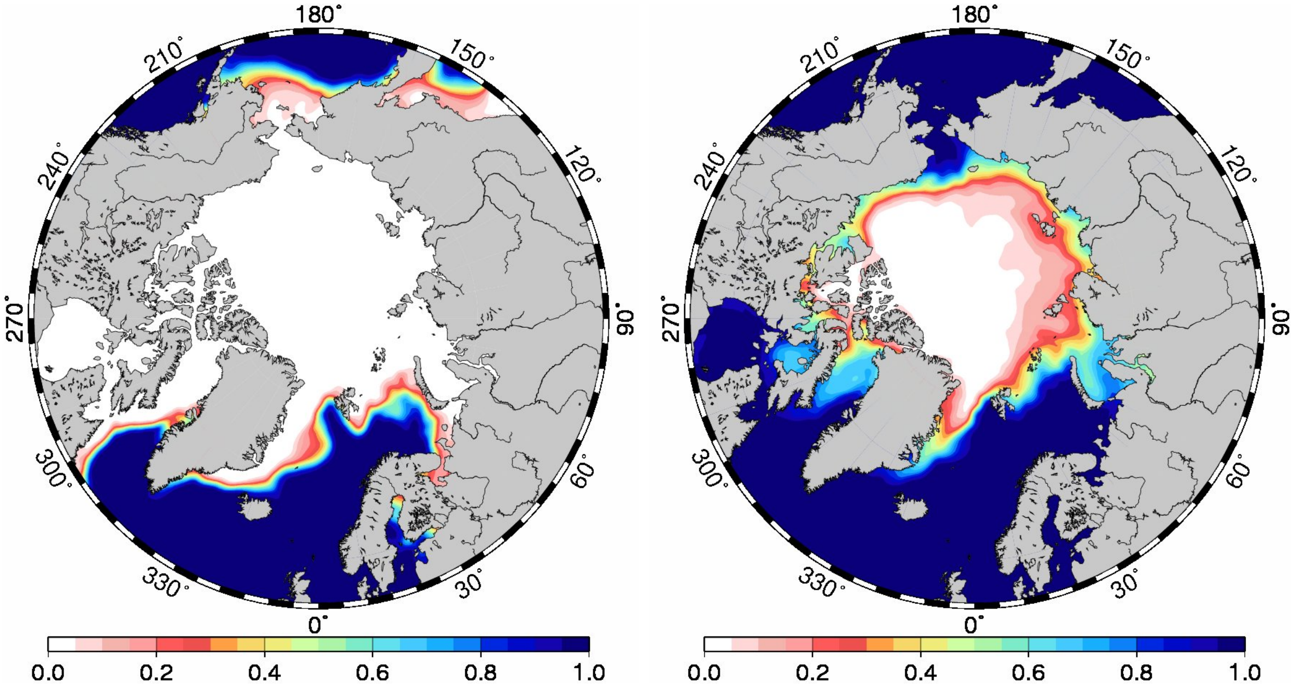

Fig. 11. Probability that the ocean is free of ice [$p(f=0)$], as computed from the non-stochastic ensemble for 15 March (left panel) and 11 September (right panel).

defined in the interval between $f=0$ (no ice) and $f=1$ (no free water).

Because of this bounded interval, it is already clear that the Gaussian model is not appropriate to describe uncertainties in ice concentrations. Moreover, the probability density function is usually maximum at one of these bounds (at $f=1$ in the middle of the ice pack, or at $f=0$ at the borders), or even at both (U-shaped pdf), which makes the Gaussian model even less appropriate. Furthermore, the two extreme values ($f=0$ or $f=1$) can often concentrate a finite probability, which means that the cdf of ice concentration makes a step at $f=0$ or $f=1$ (as explained in Sect. 2.3). Figure 11 shows for instance the probability that the ocean is free of ice ($f=0$), as computed from the ensemble for 15 March (left panel) and 11 September (right panel). In practice, the value of this probability is computed as the fraction of the ensemble members for which $f=0$. In this computation, we also applied the sample enrichment method described in Sect. 2.3, by concatenating in the local description of the probability distribution all ice concentration values in a window of 9×9 grid points. The total ensemble size at each horizontal location is thus equal to $m=81\times 748=60588$. The effect of this enrichment of the ensemble is to slightly smooth the probability maps displayed in Fig. 11, but in view of the approximations that are made in the construction of the original ensemble, there was no reason here to stay perfectly local, while the enrichment may be a good way of mitigating the inaccuracies that are related to the limited size of the available ensemble. In Fig. 11, the resulting probability increases from $p(f=0)=0$ in the interior of the ice pack, where a zero ice concentration is impossible, to $p(f=0)=1$ outside of the ice pack, where a zero ice concentration is certain (according to our assumption about the uncertainties). In the Arctic, it is also generally much larger in September (minimum ice extension) as compared to March, which shows the primary importance of resolving the seasonal cycle in the description of the probability distributions.

Strictly speaking, in presence of such probability concentrations (at $f=0$ in Fig. 11), a Gaussian anamorphosis transformation is not possible, since the cdf in Eq. (1) is not invertible. In our example, this means that several quantiles of the ensemble are equal to $f=0$, so that the piecewise linear approximation in Eq. (2) is not defined (zero denominator if $\tilde{x}_k=\tilde{x}_{k+1}$). This is why, in this example, we need to apply the approximate solution described in Sect. 2.3, which consists in modifying the quantiles of the ensemble that are equal to 0, by interpolating them between $f=0$ and the first non-zero quantile. In this particular case, this approximation amounts to replacing the Dirac at $f=0$ in the exact pdf by a boxcar function between $f=0$ and the first non-zero quantile, cumulating the same total probability as the Dirac. (Any other function to approximate the Dirac is possible by modifying the interpolation of the quantiles.) In this way, we restore the applicability of anamorphosis by transforming the non-invertible cdf into an invertible cdf, at the price of a slight spreading of the probability that is actually

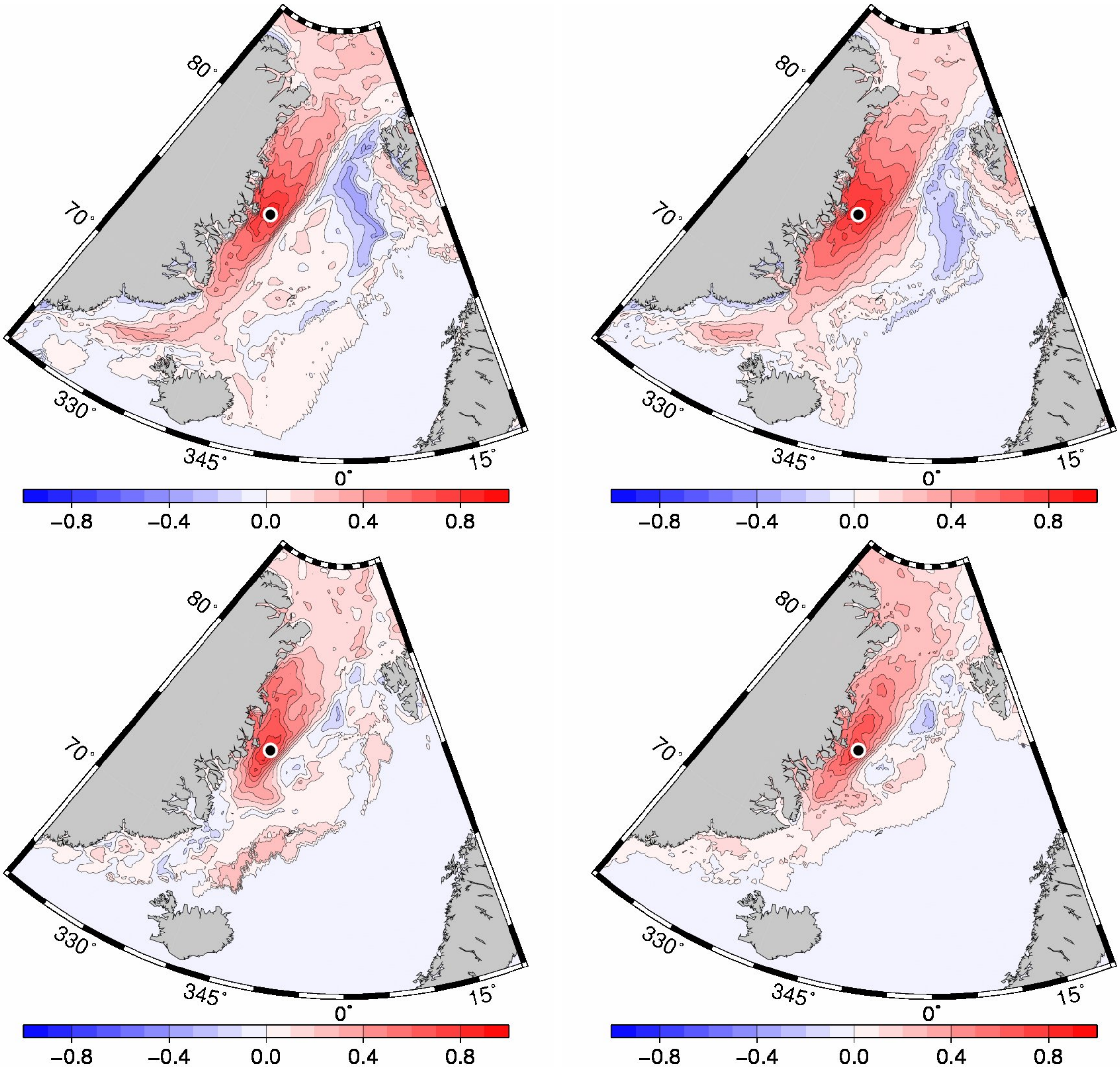

Fig. 12. Ice concentration horizontal correlation structure with respect to a reference location at 15° W 75° N (black dot) for 15 March (top panels) and 11 September (bottom panels), without anamorphosis (left panels), and after local anamorphosis transformations (right panels).

concentrated at $f = 0$. It would of course be better to avoid any kind of approximation and to keep the exact description of the probability concentrations, but this is impossible with anamorphic transformations, and it is anyway useful for data assimilation to find new variables for which the Gaussian model is (at least approximately) valid, because it makes the observational update of the prior probability distribution (with linear formulas) numerically much more efficient. And to describe the marginal probability distributions for ice concentrations, the above approximation is certainly much better than using a Gaussian model for the original variables (i.e. without anamorphic transformations).

Now, as in the previous examples, we turn to evaluating the effect of these local anamorphic transformations on the joint probability distribution by looking at the linear correlation structure. Figure 12 shows for instance the horizontal correlation structure for ice concentration with respect to a reference location at 15° W 75° N (North-East of Greenland). In the figure, we observe first that the correlation structure is very anisotropic, as a consequence of the southward ice flow along the coast of Greenland, and that the correlation distance is larger in March (Fig. 12, top panels) as compared to September (Fig. 12, bottom panels), as a result of the larger extension of the ice pack (see Fig. 11). However, in both cases, the effect of anamorphosis (in the right panels) is mainly to increase the correlation distance. In March, the correlation radius mainly increases in the cross-flow direction, because it is across the front that nonlinear dependences between the variations of ice concentrations mainly occur. And in September, the correlation radius mainly increases in the direction of the ice flow, because the reference point is then located close to the southmost edge of the ice extension.

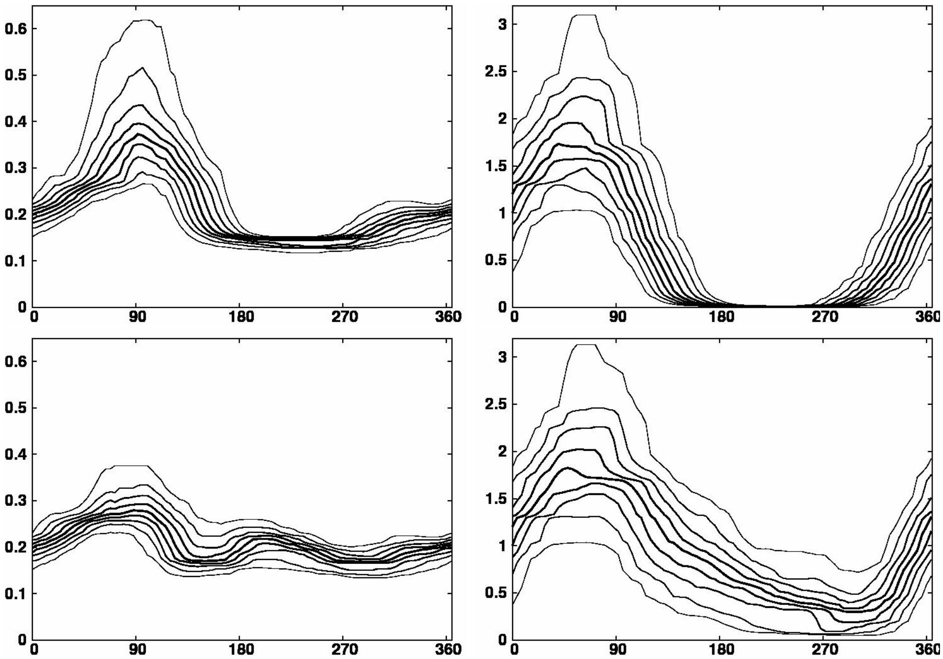

Fig. 13. Time variability of the ensemble deciles $r_k = 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9$ at 20° W 35° N (black dot in Fig. 14), as obtained for phytoplankton (left panels) and nitrate (right panels) close to the surface (top panels) and at 41 m depth (bottom panels). The further from the median ($r_k = 0.5$, thick central curve), the thinner the curve.

As a secondary effect, the anamorphosis transformations also tend to remove the spurious correlations with the exterior of the ice pack (where the probability of a zero ice concentration is close to 1). In the exterior of the ice pack, nearly all ice fractions are indeed equal to zero, so that the scatterplot with a point inside of the ice pack consists in a set of points aligned at $f = 0$, except for a few outliers, which produce the spurious correlation. The shape of the scatterplot is thus like the example 4 in Anscombe's quartet (Anscombe, 1973, Fig. 4), showing the effect of outliers on linear correlations in this typical case. By replacing the linear correlation by a nonparametric correlation, the anamorphic transformations help producing more robust correlations that are less influenced by the presence of outliers (see Sect. 2.4).

Hence, we can conclude that, in addition to significantly improving the description of the marginal probability distributions for ice concentration (in the interval between 0 and 1), local anamorphic transformations are not detrimental to the description of the horizontal correlation structure, and may even help representing nonlinear dependences between distant ice behaviours.

7 Modelling ecosystem forecast uncertainties

As a second example of non-stochastic ensemble, we study the description of ecosystem forecast uncertainties that has been used in the MyOcean project (by Fontana et al., 2012) to assimilate ocean colour data in the NEMO/LOBSTER $1/4°$ resolution CPBM (already described in Sects. 4 and 5) and produce a 9-year reanalysis (from 1998 to 2006) of the North-Atlantic ecosystem. The ensemble is constructed using the same kind of assumption as in the previous example (in Sect. 6), by sampling an interannual free model simulation (7 years, between 1999 and 2005) every 2 days in a running window of ±30 days around the date of interest (thus retaining 30 model states, every year), which makes an ensemble of size $m = 7 \times 30 = 210$ model states.

Figure 13 shows the deciles of the resulting ensemble as a function of time for phytoplankton (left panels) and nitrate (right panels) at 20° W 35° N (black dot in Fig. 14). This fully describes the approximate piecewise linear anamorphosis transformation for this location, which is defined in Eqs. (2) and (3) by a remapping of this set of deciles $\tilde{x}_k$ on the corresponding Gaussian deciles z_k. Consistently with our ensemble description of the uncertainties, only the seasonal cycle is resolved, so that the transformation is kept the same from year to year. As in the previous example, the seasonal cycle is certainly the first thing that needs to be taken into account in the description of the uncertainties. The figure indeed clearly illustrates the extreme seasonal variations in the spreading of the ensemble, in relation to the dynamics of the ecosystem. For instance, close to the surface (Fig. 13, top panels), large phytoplankton concentrations (left panel) appear during the spring bloom (around day 90), together with larger associated uncertainties. The bloom progressively depletes nitrates (right panel) until the surface concentration

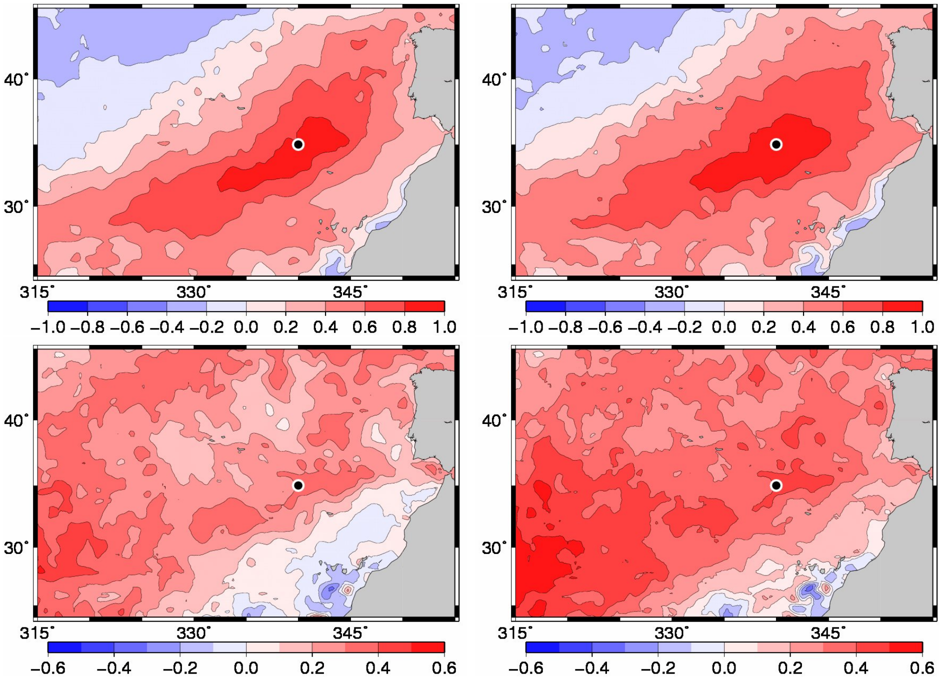

Fig. 14. Phytoplankton (top panels) and nitrate (bottom panels) horizontal correlation structure with respect to phytoplankton at 20° W 35° N (North Atlantic), without anamorphosis (left panels), and after local anamorphosis transformations (right panels).

becomes very low during the whole summer (between days 180 and 270), together with very low associated uncertainties (according to our assumption). To close the annual cycle, larger nitrate concentrations are then restored by vertical mixing during fall and winter (between days 270 and 45), when the primary production is reduced. During the whole cycle, the uncertainties on both concentrations (which are positive quantities) are clearly non-Gaussian, with the higher deciles ($r_k > 0.5$) being further away from the median than the lower deciles ($r_k < 0.5$), especially during the transitions between high and low concentrations. For instance, just before nitrates are fully depleted, the lower deciles and the median are already all close to zero, while the higher deciles are still very significant. These non-Gaussian effects are first-order behaviours of the ecosystem uncertainties, which clearly illustrate the inadequacy of the Gaussian model, and the usefulness of our approximate piecewise-linear anamorphic transformations to improve the description of the marginal probability distributions, as well as their variations in time along the annual cycle. Moreover, the dynamical characteristics of the spring bloom (amplitude, starting date,...) are known to be very heterogeneous in the ocean, so that the associated uncertainties require local transformations to be properly described. For instance, Fig. 13 (bottom panels) shows the seasonal cycle of the ensemble deciles at the same horizontal location, but at a different depth (41 m depth instead of the first model level). Here, the situation is completely changed with respect to the surface, because the spring bloom is smaller, and nitrate is not fully depleted during summer. This implies that the non-Gaussian description of the uncertainties must also be very different. See in particular the uncertainty in the nitrate concentration, which stays more symmetric around the median for the whole year. Moreover, as soon as the bloom is terminated in the surface layers (around day 180), more light becomes available at that depth, and a secondary bloom can occur during summer, together with larger phytoplankton uncertainties as compared to surface layers. The improvement in the local description of the marginal distributions already explains why the approximate anamorphosis algorithm described in Sect. 2 has been so useful in the work of Fontana et al. (2012) to improve ocean colour data assimilation.

However, it is important to check that this improvement in the description of the marginal probability distributions is not done at the expense of the joint probability distribution. And again, to evaluate if the dependence between random variables is better described by a Gaussian model before or after the anamorphosis transformations, we look at the modification of the linear correlation coefficient. Figure 14 shows for instance the PHY (top panels) and NO_3 (bottom panels) horizontal correlation structure with respect to PHY at 20° W 35° N, as obtained for the original variables (left panels) and the transformed variables (right panels), based on the ensemble obtained for 19 April, i.e. the same result as displayed in Fig. 8 for the stochastic ensemble resulting from wind random perturbations (described in Sect. 4). Concerning the

PHY correlation structure, the first thing that we observe is the same kind of anisotropy as in Fig. 8, probably reflecting some basic horizontal structure of the ecosystem dynamics, even if the correlation radius is here much larger, because the wind variability (which has been used in Sect. 4 to parameterize the statistics of wind perturbations) has a smaller decorrelation scale than the ecosystem variability in this region. But despite of this difference, the effect of anamorphosis is the same: a substantial increase of the correlation radius, especially in the direction in which the correlation radius is the smallest. This reduced anisotropy of the correlation structure after anamorphosis indicates a nonlinear dependence between the ecosystem behaviours across the frontal pattern.

Concerning the NO_3 correlation structure (Fig. 14, bottom panels), the horizontal pattern is not much changed by the local anamorphosis transformations, but the value of the cross-correlation with PHY is significantly increased. It is interesting to note that PHY and NO_3 are here positively correlated (they were anticorrelated in Fig. 8), which is the sign that, on 19 April (day 109 in Fig. 13), the short term variability dominates in the non-stochastic ensemble. This difference of behaviour between Figs. 8 and 14 can be better illustrated using scatterplots of PHY at the reference point (20° W 35° N) vs. NO_3 at some distance from the reference (20° W 33° N), as shown in Fig. 15 for the correlation structure of Fig. 8 (top panels) and Fig. 14 (bottom panels), without anamorphosis (left panels) and with anamorphosis (right panels). In the first situation (corresponding to Fig. 8), the effect of wind perturbations is to introduce more or less mixing in the water column, so that the resulting perturbation of PHY and NO_3 tend to be anticorrelated (because of their opposit vertical gradient). And in the second situation (corresponding to Fig. 14), the model variability tends to positively correlate the PHY and NO_3 fluctuations. However, in both cases, we can observe in the scatterplots that the effect of the anamorphic transformations (giving the same normalized Gaussian distribution to all marginal distributions) is to produce a scatterplot with a more elliptical shape, which is a good indication that the joint distribution is also closer to a bi-Gaussian distribution. In these cases, it can be seen that the modification of the scatterplots results from the two properties of anamorphosis that were introduced in Sect. 2: (a) the linearization of a nonlinear dependence between the two variables, and (b) the reduction of the effect of outliers (resulting here from occasional extreme behaviours). In both cases, these two properties explain the increase of linear correlation from $|\rho_{X_1X_2}| = 0.07$ to $|\rho_{Z_1Z_2}| = 0.43$ in the top panels, and from $|\rho_{X_1X_2}| = 0.24$ to $|\rho_{Z_1Z_2}| = 0.38$ in the bottom panels.

However, a closer analysis of PHY-NO_3 cross-correlations in the last example shows that they are often changing sign after the bloom event, in a way that is very heterogeneous in space and time. In addition to the improvement of the marginal distributions illustrated in Fig. 13 (in particular, the zero probability associated to negative concentrations) and to the increase of the correlation radius illustrated in Fig. 14, this ability of the scheme to adjust in space and time to local statistical behaviours is most probably one of the main reasons why it has been so helpful in the work of Fontana et al. (2012) to improve the estimate of NO_3 concentrations from ocean colour observations.

8 Conclusions

Many kinds of ocean uncertainties cannot be accurately described using a Gaussian model. This is particularly obvious in the examples of ecosystem uncertainties (in Sects. 4, 5 and 7) and sea ice uncertainties (in Sect. 6), although this may also be true for ocean dynamics uncertainties (as in the mixed layer depth example in Sect. 3). On the other hand, in these examples, a general non-Gaussian description of the joint probability distribution would be impossible to identify from a moderate size ensemble, because the uncertainties occur in too many dimensions (curse of dimensionality). Nevertheless, even with the available ensemble (a few hundred members in all examples described in the paper), it is certainly possible to go beyond the Gaussian assumption in the description of the marginal distribution for any individual random variable (including observation equivalents or indirect operational product). In this paper, we suggested that a very significant improvement can already be obtained with a very simple non-Gaussian description of the marginal distributions (histograms), based on a few quantiles of the ensemble (typically deciles, as in our examples). It is especially interesting for large size applications, because it is (i) concise (described by qn values, if n is the number of variables, and q, the number of quantiles), (ii) efficient (computational complexity proportional to $nm\log m$, if m is the size of the ensemble), and (iii) often more accurate than the Gaussian description (based on the mean and standard deviation). More importantly, this simple histogram description can also directly be used to perform a piecewise linear change of variable (anamorphosis transformation), in such a way that each marginal distribution becomes approximately Gaussian. In these transformed variables, it is then possible to perform the ensemble observational update consistently with our simple description of the marginal uncertainties, by applying the standard Gaussian algorithm, providing that the ensemble correlation structure is preserved, or even improved, by the transformation.

In the paper, various examples were used to evaluate the effect of these local anamorphic transformations on the spatial correlation structure. The results indicate that (i) the transformation is accurate enough to faithfully preserve the correlation structure if the distribution is already close to Gaussian, and (ii) the transformation has the general tendency of increasing the correlation radius as soon as the dependence between random variables becomes nonlinear. These effects may be understood by observing that the linear correlation coefficient (Pearson) between the transformed

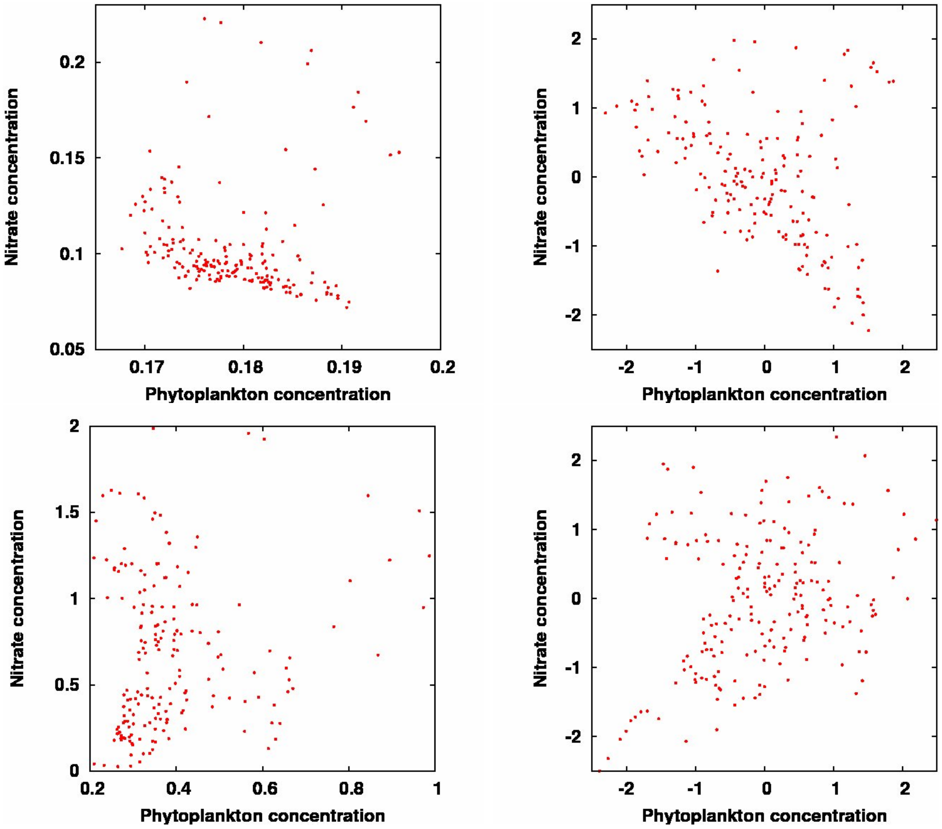

Fig. 15. Scatterplots of PHY at the reference point (20° W 35° N) vs. NO_3 at some distance from the reference (20° W 33° N), corresponding to the correlation structures that are shown in Fig. 8 (top panels) and in Fig. 14 (bottom panels), without anamorphosis (left panels) and with anamorphosis (right panels).

variables corresponds to a nonlinear measure of correlation between the original variables, which is very similar to the rank correlation (Spearman). On the other hand, even if the method finds its full justification with a stochastic ensemble description of the uncertainties, the last two examples show that it may also be useful with the non-stochastic ensembles (resulting for instance from the system past variability) that are often used in present-day operational systems to reduce the numerical cost of data assimilation (until truly stochastic solutions become affordable). In both cases, the most important consequence for data assimilation of this increase in the correlation magnitude is a significant reduction in the number of degrees of freedom in the uncertainties (in a Gaussian sense), so that a better estimation accuracy can be obtained from a given observation network. And from a more general point of view, this also means that it may sometimes be rewarding to put some time and numerical effort to improve the statistical description of the uncertainties, rather than giving too much confidence to oversimplistic assumptions.

Appendix A

Implementation issues

All examples of local anamorphic transformations described in this paper have been performed using specific tools that we have implemented in the SESAM public software[2], except the example of Sect. 6, which has been perfomed using an independent implementation of the algorithm in the Mercator assimilation system (SAM2). More specifically, the results displayed in Figs. 3 to 10, 13 and 14 have been obtained using four SESAM tools:

1. *Computation of the quantiles of the input ensemble*, with the SESAM commandline:

   ```
   sesam -mode anam  -inxbas [ens_dir]
                     -outxbasref [quant_dir]
   ```

   where *[ens_dir]* is a directory containing the input ensemble forecast (as a set of NetCDF files, using the

[2] http://www-meom.hmg.inpg.fr/SESAM

SESAM naming conventions), and *[quant_dir]*, a directory containing as an input, the definition of the quantiles (an ASCII file with the r_k, $k = 1, \ldots, q$). From this, SESAM computes the (local) quantiles of the ensemble $\tilde{x}_k$, $k = 1, \ldots, q$ (as a set of NetCDF files, in the directory *[quant_dir]*), linearly interpolating between successive ensemble members, if necessary.

2. *Local anamorphic transformation of the input ensemble*, with the SESAM commandline:

```
sesam -mode anam  -inxbas [ens_dir]
                  -inxbasref [quant_dir]
                  -outxbas [aens_dir]
                  -typeoper +
```

where *[ens_dir]* is a directory containing the input ensemble forecast, and *[quant_dir]*, a directory containing the quantiles $\tilde{x}_k$, $k = 1, \ldots, q$ of the ensemble (as obtained from the previous tool), and, as an additional input, the quantiles of the target distribution (an ASCII file with the z_k, $k = 1, \ldots, q$). From this, SESAM computes the transformed ensemble (as a set of NetCDF files, in the directory *[aens_dir]*), by linearly interpolating between the z_k using Eq. (2). In this way, the transformation can easily be performed towards any target distribution (by just changing the ASCII file with the z_k), in particular towards the Gaussian distribution (as in most examples presented in this paper) or towards the uniform distribution (using the same file for the z_k and for the r_k) as in the middle panels of Figs. 6 and 7. (The backward transformation of Eq. (3) can be performed similarly by replacing the $+$ sign by a $-$ sign in the commandline.)

3. *Computation of the EOFs of the ensemble*, with the SESAM commandline:

```
sesam -mode geof  -inxbas [(a)ens_dir]
                  -outxbas [(a)eof_dir]
```

where *[(a)ens_dir]* is a directory containing the input or transformed ensemble, from which SESAM computes the EOFs (as a set of NetCDF files, in the directory *[(a)eof_dir]*). This tool may be useful to obtain an orthogonal basis of the linear subspace spanned by the (original or transformed) ensemble forecast, or to reduce the rank of the ensemble covariance matrix (by discarding the directions with negligible variance). No rank reduction has been performed in the examples described in this paper.

4. *Computation of the correlation structure*, with the SESAM commandline:

```
sesam -mode corr  -inxbas [(a)eof_dir]
                  -outvar [corr_file]
                  -incfg [cfg_file]
```

where *[(a)eof_dir]* is a directory containing the EOFs of the original or transformed ensemble (or the columns of any other square root of the ensemble covariance matrix), and *[cfg_file]* is a configuration file describing the reference variable (an ASCII file, with the name of the variable, and the grid coordinates). From this, SESAM computes the multivariate correlation structure with respect to the reference variable (as a NetCDF file *[corr_file]* providing the corresponding column of the correlation matrix). This is the kind of result that is mostly displayed throughout this paper.

Hence, only four SESAM commandlines have been sufficient to produce all kinds of result that have been presented in this paper, for a variety of oceanographic applications. The first one (1) provides the histogram description of the marginal uncertainties. This is used by the second one (2) to perform the piecewise linear local anamorphic transformation, as a preprocessing to any operation taking profit from Gaussianity, like the computation of EOFs (3), the diagnostic of the linear correlation structure (4) or the linear observational update (not shown here). In this way, the same study can be easily repeated to any new oceanographic problem, to check if the same conclusions apply. In our view, the simplicity and modularity of the implementation is an additional argument speaking in favour of the approximate algorithm described in Sect. 2.

Acknowledgements. This work was conducted as a contribution to the MyOcean and SANGOMA projects funded by the EU (grants FP7-SPACE-2007-1-CT-218812-MYOCEAN and FP7-SPACE-2011-1-CT-283580-SANGOMA), with additional support from the ASSOCO project (ESA/ESRIN Contract Network 22408/09/I-EC) for M. Doron and CNES for M. Meinvielle. The calculations were performed using HPC resources from GENCI-IDRIS (Grant 2010-011279).

Edited by: J. Schröter

cnrs INSU Institut national des sciences de l'Univers

The publication of this article is financed by CNRS-INSU.

References

Anscombe, F. J.: Graphs in Statistical Analysis, American Statistician, 27, 17–21, 1973.

Béal, D., Brasseur, P., Brankart, J.-M., Ourmières, Y., and Verron, J.: Characterization of mixing errors in a coupled physical biogeochemical model of the North Atlantic: implications for nonlinear estimation using Gaussian anamorphosis, Ocean Sci., 6, 247–262, doi:10.5194/os-6-247-2010, 2010.

Bertino, L., Evensen, G., and Wackernagel, H.: Sequential Data Assimilation Techniques in Oceanography, Int. Statist. Rev., 71, 223–241, 2003.

Bocquet, M., Pires, C. A., and Wu, L.: Beyond Gaussian statistical modeling in geophysical data assimilation, Mon. Weather Rev., 138, 2997–3023, 2010.

Chilès, J.-P. and Delfiner, P.: Geostatistics: Modeling Spatial Uncertainty, Wiley, New York, 1999.

Corder, G. W. and Foreman, D. I.: Nonparametric Statistics for Non-Statisticians: A Step-by-Step Approach, Wiley, 2009.

Cover, T. M. and Thomas, J. A.: Elements of information theory, Wiley, 2006.

Doron, M., Brasseur, P., and Brankart, J.-M.: Estimation of biogeochemical parameters of a 3D ocean coupled physical-biogeochemical model with a stochastic data assimilation method: twin experiments., J. Mar. Syst., 87, 194–207, 2011.

Evensen, G. and van Leeuwen, P. J.: Assimilation of Geosat Altimeter Data for the Agulhas Current using the Ensemble Kalman Filter with a Quasi-Geostrophic Model, Mon. Weather Rev., 124, 85–96, 1996.

Fontana, C., Brasseur, P., and Brankart, J.-M.: Toward a multivariate reanalysis of the North Atlantic ocean biochemistry during 1998–2006 based on the assimilation of SeaWiFS chlorophyll data, Ocean Sci. Discuss., in preparation, 2012.

Hollander, M. and Wolfe, D. A.: Nonparametric statistical methods, Wiley, 1973.

Izenman, A. J.: Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning, Springer, 2008.

Kendall, M. G.: Rank correlation methods, Griffin, 1962.

Lauvernet, C., Brankart, J.-M., Castruccio, F., Broquet, G., Brasseur, P., and Verron, J.: A truncated Gaussian filter for data assimilation with inequality constraints: application to the hydrostatic stability condition in ocean models, Ocean Modeling, 27, 1–17, 2009.

Lermusiaux, P. F. J.: Uncertainty Estimation and Prediction for Interdisciplinary Ocean Dynamics, J. Comput. Phys., 217, 176–199, 2006.

Lévy, M., Gavart, M., Mémery, L., Caniaux, G., and Paci, A.: A four-dimensional mesoscale map of the spring bloom in the northeast Atlantic (POMME experiment): results of a prognostic model, J. Geophys. Res., 110, C07S21, doi:10.1029/2004JC002588, 2005.

Longhurst, A.: Seasonal cycles of pelagic production and consumption., Prog. Oceanogr., 32, 77–167, 1995.

Madec, G. and Imbard, M.: A global ocean mesh to overcome the North Pole singularity, Clim. Dynam., 12, 381–388, 1996.

Meinvielle, M.: Ajustement optimal des paramètres de forçage atmosphérique par assimilation de données de température de surface pour des simulations océaniques globales, Ph.D. thesis, Université Joseph Fourier (Grenoble), 2011.

Ourmières, Y., Brasseur, P., Lévy, M., Brankart, J.-M., and Verron, J.: On the key role of nutrient data to constrain a coupled physical-biogeochemical assimilative model of the North Atlantic Ocean, J. Mar. Syst., 75, 100–115, 2009.

Palmer, T. N., Shutts, G. J., Hagedorn, R., Doblas-Reyes, F. J., Jung, T., and Leutbecher, M.: Representing model uncertainty in weather and climate prediction, Annu. Rev. Earth Planet. Sci., 33, 163–193, 2005.

Silverman, B. W.: Density Estimation for Statistics and Data Analysis, Chapman and Hall, 1986.

Simon, E. and Bertino, L.: Application of the Gaussian anamorphosis to assimilation in a 3-D coupled physical-ecosystem model of the North Atlantic with the EnKF: a twin experiment, Ocean Sci., 5, 495–510, doi:10.5194/os-5-495-2009, 2009.

Skandrani, C., Brankart, J.-M., Ferry, N., Verron, J., Brasseur, P., and Barnier, B.: Controlling atmospheric forcing parameters of global ocean models: sequential assimilation of sea surface Mercator-Ocean reanalysis data, Ocean Sci., 5, 403–419, doi:10.5194/os-5-403-2009, 2009.

Spearman, C.: The proof and measurement of association between two things, Amer. J. Psychol., 15, 72–101, 1904.

The DRAKKAR Group: Eddy-permitting Ocean Circulation Hindcasts of past decades, CLIVAR Exchanges 42, 12, 8–10, 2007.

van Leeuwen, P.-J.: Particle filtering in geophysical systems, Mon. Weather Rev., 137, 4089–4114, 2009.

van Leeuwen, P.-J.: Nonlinear Data Assimilation in geosciences: an extremely efficient particle filter, Q. J. Roy. Meteorol. Soc., 136, 1991–1996, 2010.

Von Mises, R.: Mathematical Theory of Probability and Statistics, Academic Press, New York, 1964.

Wackernagel, H.: Multivariate Geostatistics, Springer, 2003.