Manuscript prepared for Nonlin. Processes Geophys. Discuss.
with version 2015/04/24 7.83 Copernicus papers of the LaTeX class copernicus.cls.
Date: 28 September 2015

Expanding the validity of the ensemble Kalman filter without the intrinsic need for inflation

Marc Bocquet[1], **Patrick Nima Raanes**[2,3], **and Alexis Hannart**[4]

[1]CEREA, Joint laboratory École des Ponts ParisTech and EDF R&D, Université Paris-Est, Champs-sur-Marne, France
[2]Nansen Environmental and Remote Sensing Center, Bergen, Norway
[3]Mathematical Institute, University of Oxford, United-Kingdom
[4]IFAECI, CNRS-CONICET-UBA, Buenos Aires, Argentina

Correspondence to: Marc Bocquet
(bocquet@cerea.enpc.fr)

Abstract

The ensemble Kalman filter (EnKF) is a powerful data assimilation method meant for high-dimensional nonlinear systems. But its implementation requires fixes such as localization and inflation. The recently developed *finite-size* ensemble Kalman filter (EnKF-N) does not require multiplicative inflation meant to counteract sampling errors. Aside from the practical interest of avoiding the tuning of inflation in perfect model data assimilation experiments, it also offers theoretical insights and a unique perspective on the EnKF. Here, we revisit, clarify and correct several key points of the EnKF-N derivation. This simplifies the use of the method, and expands its validity. The EnKF is shown to not only rely on the observations and the forecast ensemble but also on an implicit prior assumption, termed *hyperprior*, that fills in the gap of missing information. In the EnKF-N framework, this assumption is made explicit through a Bayesian hierarchy. This hyperprior has been so far chosen to be the uninformative *Jeffreys*' prior. Here, this choice is revisited to improve the performance of the EnKF-N in the regime where the analysis is strongly dominated by the prior. Moreover, it is shown that the EnKF-N can be extended with a normal-inverse-Wishart informative hyperprior that introduces additional information on error statistics. This can be identified as a hybrid EnKF-3D-Var counterpart to the EnKF-N.

1 Introduction

The ensemble Kalman filter (EnKF) has become a popular data assimilation method for high-dimensional geophysical systems (**?**, and references therein). The flow-dependence of the forecast error used in the analysis is its main strength, compared to schemes using static background statistics such as 3D-Var and 4D-Var.

However, to perform satisfyingly, the EnKF may require the use or inflation and/or localization, depending on the data assimilation system setup. Localization is required in the rank-deficient regime, in which the limited size of the ensemble leads to an empirical error covariance matrix of too small rank, as is often the case in realistic high-dimensional

systems (**???**). It can also be useful in a rank-sufficient context in the presence of non-Gaussian/non-linear effects.

Inflation is a complementary technique meant to increase the variances diagnosed by the EnKF (**??**). It is usually intended to compensate for an underestimation of uncertainty. This underestimation can be caused either by sampling error, an intrinsic deficiency of the EnKF system, or model error, an extrinsic deficiency.

A variant of the EnKF, called the finite-size ensemble Kalman filter (EnKF-N) has been introduced in **??**. It has subsequently been successfully applied in **??** in an ensemble variational context. It has been shown to avoid the need for multiplicative inflation usually needed to counteract sampling errors. In particular, it avoids the costly chore of tuning this inflation.

The EnKF-N is derived by assuming that the ensemble members are drawn from the same distribution as the truth, but makes no further assumptions on the ensemble's accuracy. In particular, the EnKF-N, unlike the traditional EnKFs, does not make the approximation that the sample first- and second-order moments coincide with the actual moments of the prior (which would be accessible if the ensemble size N was infinite).

Through its mathematical derivation, the scheme underlines the missing information besides the observations and the ensemble forecast, an issue which is ignored by traditional EnKFs. This missing information is explicitly compensated for in the EnKF-N using a so-called *hyperprior*. In **?**, a simple choice was made for this hyperprior, namely the Jeffreys' prior, which is meant to be as much non-informative as possible. While the EnKF-N built on Jeffreys' prior often performs very well with low-order models, it may fail in specific dynamical regimes because a finer hyperprior is needed (**?**). Other choices were made in the derivation of the EnKF-N which remain only partly justified or insufficiently clear.

The objective of this paper is to clarify several of those choices, to answer several questions raised in the above references, and to advocate the use of improved or new hyperpriors. This should add to the theoretical understanding of the EnKF, but also provide a useful algorithm. Specifically, the EnKF-N allows the development of data assimilation systems under perfect model conditions without worrying about tuning the inflation. In the whole paper, we will restrict ourselves to perfect model conditions.

In Section 2, the key ideas and algorithms of the EnKF-N are recalled and several aspects of the approach are clarified. It is shown that the redundancy in the EnKF centered perturbations leads to a subtle but important correction to the EnKF-N when the analysis is performed in the affine space defined by the mean state and the ensemble perturbations. In Section 3, the ensemble update step of the EnKF-N is revisited and clarified. In Section 4, the nonlinearity of the ensemble forecast step and its handling by the EnKF-N, and more generally multiplicative inflation, are discussed. The corrections to the EnKF-N are illustrated with numerical experiments in Section 5. Sections 6 and 7 discuss of modifying or even changing the hyperprior. In Section 6, we discuss caveats of the method in regimes where the posterior ensemble is drawn to the prior ensemble. Simple alternatives to the Jeffreys' hyperprior are proposed. Finally, a class of more informative priors based on the normal-inverse-Wishart distribution and permitting to incorporate additional information on error statistics is introduced and theoretically discussed in Section 7. Conclusions are given in Section 8.

2 The finite-size ensemble Kalman filter (EnKF-N)

The key ideas of the EnKF-N are presented and clarified in this section. Additional insights into the scheme and why it is successful are also given.

2.1 Marginalizing over potential priors

? (later **?**) recognized that the ensemble mean $\overline{\mathbf{x}}$ and ensemble error covariance matrix $\mathbf{P}$ used in the EnKF may be different from the unknown first- and second-order moments of the true error distribution, $\mathbf{x}_b$ and $\mathbf{B}$, where $\mathbf{B}$ is a positive definite matrix. The mismatch is due to the finite-size of the ensemble which leads to sampling errors, partially induced by the nonlinear ensemble propagation in the forecast step (see Section 4). Figure 1 illustrates the effect of sampling error when the prior is assumed Gaussian and reliable, whereas the prior actually stems from an uncertain sampling using the ensemble.

The EnKF-N prior accounts for the uncertainty in $\mathbf{x}_b$ and $\mathbf{B}$. Denote $\mathbf{E} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N]$ the ensemble of size N formatted as an $M \times N$ matrix where M is the state space dimension, $\overline{\mathbf{x}} = \mathbf{E}\mathbf{1}/N$ the ensemble mean where $\mathbf{1} = (1, \cdots, 1)^{\mathsf{T}}$, and $\mathbf{X} = \mathbf{E} - \overline{\mathbf{x}}\mathbf{1}^{\mathsf{T}}$ the perturbation matrix. Hence, $\mathbf{P} = \mathbf{X}\mathbf{X}^{\mathsf{T}}/(N-1)$ is the empirical covariance matrix of the ensemble. Marginalizing over all potential $\mathbf{x}_b$ and $\mathbf{B}$, the prior of $\mathbf{x}$ reads

$$p(\mathbf{x}|\mathbf{E}) = \int \mathrm{d}\mathbf{x}_b \mathrm{d}\mathbf{B}\, p(\mathbf{x}|\mathbf{E}, \mathbf{x}_b, \mathbf{B}) p(\mathbf{x}_b, \mathbf{B}|\mathbf{E})\,. \tag{1}$$

The symbol $\mathrm{d}\mathbf{B}$ corresponds to the Lebesgue measure on all independent entries $\prod_{i \leq j}^{M} \mathrm{d}[\mathbf{B}]_{ij}$, but the integration is restricted to the cone of positive definite matrices. Since $p(\mathbf{x}|\mathbf{E}, \mathbf{x}_b, \mathbf{B})$ is conditioned on the knowledge of the true prior statistics and assumed to be Gaussian, it does not depend on $\mathbf{E}$, so that:

$$p(\mathbf{x}|\mathbf{E}) = \int \mathrm{d}\mathbf{x}_b \mathrm{d}\mathbf{B}\, p(\mathbf{x}|\mathbf{x}_b, \mathbf{B}) p(\mathbf{x}_b, \mathbf{B}|\mathbf{E})\,. \tag{2}$$

Bayes' rule can be applied to $p(\mathbf{x}_b, \mathbf{B}|\mathbf{E})$, yielding

$$p(\mathbf{x}|\mathbf{E}) = \frac{1}{p(\mathbf{E})} \int \mathrm{d}\mathbf{x}_b \mathrm{d}\mathbf{B}\, p(\mathbf{x}|\mathbf{x}_b, \mathbf{B}) p(\mathbf{E}|\mathbf{x}_b, \mathbf{B}) p(\mathbf{x}_b, \mathbf{B})\,. \tag{3}$$

Assuming independence of the samples, the likelihood of the ensemble $\mathbf{E}$ can be written

$$p(\mathbf{E}|\mathbf{x}_b, \mathbf{B}) = \prod_{n=1}^{N} p(\mathbf{x}_n|\mathbf{x}_b, \mathbf{B})\,. \tag{4}$$

The last factor, $p(\mathbf{x}_b, \mathbf{B})$, is the hyperprior. This distribution represents our beliefs about the forecasted filter statistics, $\mathbf{x}_b$ and $\mathbf{B}$, prior to actually running any filter. This distribution is termed hyperprior because it represents a prior for the background information in the first stage of a Bayesian hierarchy.

Assuming one subscribes to this EnKF-N view on the EnKF, it shows that additional information is actually required in the EnKF, in additional to the observations and the prior ensemble which are potentially insufficient to make an inference.

A simple choice was made in **?** for the hyperprior: the Jeffreys' prior is an analytically tractable and uninformative hyperprior of the form

$$p_{\mathrm{J}}(\mathbf{x}_b, \mathbf{B}) \propto |\mathbf{B}|^{-\frac{M+1}{2}} , \tag{5}$$

where $|\mathbf{B}|$ is the determinant of the background error covariance matrix $\mathbf{B}$ of dimension $M \times M$.

2.2 Predictive prior

With a given hyperprior, the marginalization over $\mathbf{x}_b$ and $\mathbf{B}$, Eq. (3), can in principle be carried out to obtain $p(\mathbf{x}|\mathbf{E})$. We choose to call it a *predictive prior* to comply with the traditional view that sees it as prior before assimilating the observations. Note, however, that statisticians would rather call it a *predictive posterior* distribution as the outcome of a first-stage inference of a Bayesian hierarchy, where $\mathbf{E}$ is the data.

Using Jeffreys' hyperprior, **?** showed that the integral can be obtained analytically and that the predictive prior is a multivariate T-distribution:

$$p(\mathbf{x}|\mathbf{E}) \propto \left| \frac{(\mathbf{x} - \overline{\mathbf{x}})(\mathbf{x} - \overline{\mathbf{x}})^{\mathsf{T}}}{N-1} + \varepsilon_N \mathbf{P} \right|^{-\frac{N}{2}} , \tag{6}$$

where $|.|$ denotes the determinant and $\varepsilon_N = 1 + 1/N$. The determinant is computed in the space generated by the perturbations of the ensemble so that it is not singular. This distribution has fat tails thus accounting for the uncertainty in $\mathbf{B}$. The factor ε_N is a result of the uncertainty in $\mathbf{x}_b$; if $\mathbf{x}_b$ were known to coincide with the ensemble mean $\overline{\mathbf{x}}$, then ε_N would be 1 instead. For a Gaussian process, $\varepsilon_N \mathbf{P}$ is an unbiased estimator of the squared error of the ensemble mean $\overline{\mathbf{x}}$ (**?**), where ε_N stems from the uncertain $\mathbf{x}_b$ which does not coincide with

$\overline{\mathbf{x}}$. In the derivation of **?**, the $\varepsilon_N \mathbf{P}$ correction comes from integrating out on $\mathbf{x}_b$. Therefore, ε_N can be seen as an inflation factor on the prior covariance matrix that should actually apply to any type of EnKF.

This non-Gaussian prior distribution can be seen as an average over Gaussian distributions weighted according to the hyperprior. It can be shown that Eq. (6) can be re-arranged:

$$p(\mathbf{x}|\mathbf{E}) \propto \left\{ 1 + \frac{(\mathbf{x}-\overline{\mathbf{x}})^{\mathsf{T}} (\varepsilon_N \mathbf{P})^{\dagger} (\mathbf{x}-\overline{\mathbf{x}})}{N-1} \right\}^{-\frac{N}{2}}, \tag{7}$$

where $\mathbf{P}^{\dagger}$ is the Moore-Penrose inverse of $\mathbf{P}$.

In comparison, the traditional EnKF implicitly assumes that the hyperprior is $\delta(\mathbf{B}-\mathbf{P})\delta(\mathbf{x}_b - \overline{\mathbf{x}})$ where δ is a Dirac multidimensional distribution. In other words the background statistics generated from the ensemble coincide with the true background statistics. As a result, one obtains in this case the Gaussian prior:

$$p(\mathbf{x}|\mathbf{E}) \propto \exp\left\{ -\frac{1}{2} (\mathbf{x}-\overline{\mathbf{x}})^{\mathsf{T}} \mathbf{P}^{\dagger} (\mathbf{x}-\overline{\mathbf{x}}) \right\}. \tag{8}$$

2.3 Analysis

Consider a given analysis step of the data assimilation cycle. The observation vector is denoted $\mathbf{y}$ of dimension d. In a Bayesian analysis, $p(\mathbf{x}|\mathbf{y}) = p(\mathbf{y}|\mathbf{x})p(\mathbf{x})/p(\mathbf{y})$, the likelihood $p(\mathbf{y}|\mathbf{x})$ is decoupled from the prior pdf $p(\mathbf{x})$. In the EnKF-N framework we are interested in $p(\mathbf{x}|\mathbf{y},\mathbf{E})$. Bayes' formula then reads

$$p(\mathbf{x}|\mathbf{y},\mathbf{E}) = \frac{p(\mathbf{y}|\mathbf{x},\mathbf{E})p(\mathbf{x}|\mathbf{E})}{p(\mathbf{y}|\mathbf{E})}. \tag{9}$$

But $\mathbf{y}$ does not depend on $\mathbf{E}$ when conditioned on $\mathbf{x}$: $p(\mathbf{y}|\mathbf{x},\mathbf{E}) = p(\mathbf{y}|\mathbf{x})$. As a consequence, Bayes' formula now simply reads within the EnKF-N framework:

$$p(\mathbf{x}|\mathbf{y},\mathbf{E}) = \frac{p(\mathbf{y}|\mathbf{x})p(\mathbf{x}|\mathbf{E})}{p(\mathbf{y}|\mathbf{E})}. \tag{10}$$

This is at odds with the ill-founded claim by **?** that the likelihood still depends on **E**. This expression clarifies one of the issue raised in **?**.

Let us recall and further discuss the analysis step of the EnKF-N for state estimation. For the sake of simplicity, the observational error distribution is assumed Gaussian, unbiased, with covariance matrix $\mathbf{R}$. The observation operator will be denoted H. Because the predictive prior Eq. (6) is non-Gaussian, the analysis is performed through a variational optimization similarly to the maximum likelihood filter (**?**), rather than by matrix algebra as in traditional EnKFs. Working in ensemble space, states are parameterized by vectors $\mathbf{w}$ of size N such that

$$\mathbf{x} = \overline{\mathbf{x}} + \mathbf{X}\mathbf{w} \,. \tag{11}$$

There is at least one "gauge" degree of freedom in $\mathbf{w}$ due to the fact that $\mathbf{x}$ is invariant under $\mathbf{w} \mapsto \mathbf{w} + \lambda \mathbf{1}$, where λ is an arbitrary scalar. This is the result of the linear dependence of the centered perturbation vectors.

For reference, with these notations, the cost function of the ensemble transform Kalman filter (ETKF, **??**) based on Eq. (8) reads:

$$\mathcal{J}(\mathbf{w}) = \frac{1}{2} \left\| \mathbf{y} - H(\overline{\mathbf{x}} + \mathbf{X}\mathbf{w}) \right\|^2_{\mathbf{R}} + \frac{N-1}{2} \left\| \mathbf{w} \right\|^2_{\mathbf{\Pi_w}} \tag{12}$$

where $\|\mathbf{z}\|^2_{\mathbf{R}} = \mathbf{z}^{\mathsf{T}} \mathbf{R}^{-1} \mathbf{z}$ and $\mathbf{\Pi_w}$ is the orthogonal projector onto the row space of $\mathbf{X}$. Algebraically, $\Pi_{\mathbf{w}} = \mathbf{X}^{\dagger}\mathbf{X}$ where $\mathbf{X}^{\dagger}$ is the Moore-Penrose inverse of $\mathbf{X}$. Equation (12) is the direct result of the substitution into Eq. (8) of $\mathbf{x}$ by $\mathbf{w}$ using Eq. (11). As explained by **?**, one can add the term $\|\mathbf{w}\|^2_{\mathbf{I}_N - \mathbf{\Pi_w}}$ to the cost function without altering the minimum. Denoting $\|\mathbf{z}\|^2 = \mathbf{z}^{\mathsf{T}}\mathbf{z}$, this leads to:

$$\mathcal{J}(\mathbf{w}) = \frac{1}{2} \left\| \mathbf{y} - H(\overline{\mathbf{x}} + \mathbf{X}\mathbf{w}) \right\|^2_{\mathbf{R}} + \frac{N-1}{2} \left\| \mathbf{w} \right\|^2 . \tag{13}$$

The added term has been labelled gauge fixing term by **?** using standard physics terminology. The EnKF-N cost function in **?** is

$$\mathcal{J}(\mathbf{w}) = \frac{1}{2} \left\| \mathbf{y} - H(\overline{\mathbf{x}} + \mathbf{X}\mathbf{w}) \right\|^2_{\mathbf{R}} + \frac{N}{2} \ln\left(\varepsilon_N + \left\| \mathbf{w} \right\|^2 \right) . \tag{14}$$

It is the result of the substitution of $\mathbf{x}$ by $\mathbf{w}$ using Eq. (11) into Eq. (7), and of the addition of the gauge fixing term albeit inside the logarithm, which was justified by extending the idea of **?** and the monotonicity of the logarithm. The restriction of $\mathbf{x}$ to the ensemble subspace is an approximation inherent in the traditional EnKFs. By virtue of the hyperprior, it is not necessarily part of the EnKF-N. However, it is quite justified assuming the ensemble tracks the unstable subspace of the dynamics.

There is another caveat in the use of the ensemble transform Eq. (11). First of all, the logarithm of the determinant of the Jacobian matrix should be added to the cost function since

$$\ln p_{\mathbf{w}}(\mathbf{w}) = \ln p_{\mathbf{x}}(\mathbf{x}(\mathbf{w})) + \ln \left| \frac{\partial \mathbf{x}(\mathbf{w})}{\partial \mathbf{w}} \right|. \tag{15}$$

Had the transformation $\mathbf{w} \mapsto \mathbf{x}(\mathbf{w})$ been nonlinear, the cost function would have been impacted (see for instance **?**). However, the standard ensemble transform is linear which should result in an irrelevant constant. Unfortunately, because of the gauge degree(s) of freedom of the perturbations, the transformation is not injective and therefore singular, and the determinant of the transformation is zero yielding an undefined constant. Hence, the issue should be addressed more carefully. It will turn out in the following section that the cost function should be amended in the non-quadratic case.

2.4 Accounting for the gauge degrees of freedom of the ensemble transform

Let us denote $\widetilde{N} \leq \min(N-1, M)$ the rank of $\mathbf{X}$. The number of gauge degrees of freedom is then $g \equiv N - \widetilde{N}$. The most common case encountered when applying the EnKF to high-dimensional systems is that the rank of $\mathbf{X}$ is $N - 1 \ll M$, that is to say $g = 1$ because $\mathbf{X}\mathbf{1} = \mathbf{0}$. A non singular ensemble transform is obtained by restricting $\mathbf{w}$ to $\mathcal{N}^\perp$ the orthogonal complement of the null space, $\mathcal{N}$, of $\mathbf{X}$. Hence, the ensemble transform:

$$\begin{aligned} T: \quad \mathcal{N}^\perp \quad &\longrightarrow T(\mathcal{N}^\perp) \\ \widetilde{\mathbf{w}} \quad &\longmapsto T(\widetilde{\mathbf{w}}) = \mathbf{X}\widetilde{\mathbf{w}} \end{aligned} \tag{16}$$

is nonsingular. This amounts to fixing the gauge at zero. With this restriction to $\mathcal{N}^\perp$, the prior of the ETKF defined over $\mathcal{N}^\perp$ is

$$p(\widetilde{\mathbf{w}}) \propto \exp\left(-\frac{N-1}{2}\|\widetilde{\mathbf{w}}\|^2\right), \tag{17}$$

whereas the prior pdf of the EnKF-N is

$$p(\widetilde{\mathbf{w}}) \propto \left(\varepsilon_N + \|\widetilde{\mathbf{w}}\|^2\right)^{-\frac{N}{2}}. \tag{18}$$

In principle, the analysis can be performed in $\mathcal{N}^\perp$ using reduced variables $\mathbf{w}_r \in \mathbb{R}^{\tilde{N}}$, looking for an estimate of the form $\mathbf{x} = \overline{\mathbf{x}} + \mathbf{X}_r\mathbf{w}_r$, where $\mathbf{X}_r$ would stand for a reduced perturbation matrix. To do so, let us introduce the singular value decomposition of the initial perturbation matrix: $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\mathsf{T}$, with $\mathbf{U} \in \mathbb{R}^{M\times\tilde{N}}$ such that $\mathbf{U}^\mathsf{T}\mathbf{U} = \mathbf{I}_{\tilde{N}}$, $\mathbf{\Sigma}$ is a diagonal positive matrix in $\mathbb{R}^{\tilde{N}^2}$, and $\mathbf{V} \in \mathbb{R}^{N\times\tilde{N}}$ is such that $\mathbf{V}^\mathsf{T}\mathbf{V} = \mathbf{I}_{\tilde{N}}$. The reduced perturbation matrix $\mathbf{X}_r$ is then simply given by $\mathbf{X}_r = \mathbf{U}\mathbf{\Sigma}$. However, the change of variable $\mathbf{w} \mapsto \mathbf{w}_r$ would prevent us from using the elegant symmetric formalism of the ensemble transform Kalman filter because the perturbation matrix $\mathbf{X}_r$ is not centered. Moreover, the new perturbations, $\mathbf{X}_r$, are non-trivial linear combinations of the initial perturbations, $\mathbf{X}$. It is likely to generate imbalances with nonlinear dynamics. Indeed, it is unlikely that the displacement of the ensemble in the analysis would be minimized, as opposed to what happens with the ETKF when the transform matrix is chosen symmetric (**?**). We applied this change of variable to a standard ETKF and tested it numerically with the Lorenz-95 low-order model (**?**). We obtained much larger displacements and intermittent instabilities that require more inflation.

Hence, we wish to fix the gauge while keeping the initial perturbations as much as possible. To do so, the definition of the prior pdfs defined on $\mathcal{N}^\perp$ are extended to the full ensemble space $\mathbb{R}^N = \mathcal{N}^\perp \oplus \mathcal{N}$, while maintaining their correct marginal over $\mathcal{N}^\perp$. For the EnKF, we can fix the gauge by choosing

$$p(\mathbf{w}) \propto \exp\left(-\frac{N-1}{2}\|\mathbf{w}\|^2\right), \tag{19}$$

as in Eq. (13) which has indeed the correct marginal since $p(\mathbf{w})$ factorizes into independent components for $\mathcal{N}$ and $\mathcal{N}^\perp$. For the EnKF-N, we can fix the gauge while keeping the symmetry by choosing

$$p(\mathbf{w}) \propto \left(\varepsilon_N + \|\mathbf{w}\|^2\right)^{-\frac{N+g}{2}} . \tag{20}$$

It can be seen that this pdf has the correct marginal by integrating out on $\mathcal{N}$, using the change of variable $\mathbf{w} - \widetilde{\mathbf{w}} \mapsto \sqrt{\varepsilon_N + \|\widetilde{\mathbf{w}}\|^2}(\mathbf{w} - \widetilde{\mathbf{w}})$.

The use of these extended pdfs in the analysis are justified by the fact that the Bayesian analysis pdf $p(\mathbf{w}|\mathbf{y})$ in ensemble space has the correct marginal over $\mathcal{N}^\perp$. Indeed, if $p(\mathbf{y}|\mathbf{w}) = p(\mathbf{y}|\mathbf{x} = \overline{\mathbf{x}} + \mathbf{X}\mathbf{w})$ is the likelihood in ensemble space which does not depend on $\widetilde{\mathbf{w}}$, then the marginal of the Bayesian analysis pdf $p(\mathbf{w}|\mathbf{y}) \propto p(\mathbf{y}|\mathbf{w})p(\mathbf{w})$ is consistently given by $p(\widetilde{\mathbf{w}}|\mathbf{y}) \propto p(\mathbf{y}|\widetilde{\mathbf{w}})p(\widetilde{\mathbf{w}})$. We conclude that it is possible to perform an analysis in terms of the redundant $\mathbf{w}$ in place of $\widetilde{\mathbf{w}}$.

As opposed to the Gaussian case, the form of pdf Eq. (20) brings in a change in the EnKF-N when the analysis is performed in ensemble space. The appearance of g in the exponent is due to a non trivial Jacobian determinant when passing from the ungauged to the gauged variables, a minimalist example of the so-called Faddeev-Popov determinant (**?**). This consideration generates a modification of the EnKF-N cost function when using Eq. (20) as the predictive prior. Henceforth, we shall assume $g = 1$, which will always be encountered in the rest of the paper. Consequently, the modified EnKF-N has the following cost function:

$$\mathcal{J}(\mathbf{w}) = \frac{1}{2}\|\mathbf{y} - H(\overline{\mathbf{x}} + \mathbf{X}\mathbf{w})\|^2_{\mathbf{R}} + \frac{N+1}{2}\ln\left(\varepsilon_N + \|\mathbf{w}\|^2\right), \tag{21}$$

which replaces Eq. (14). This modification, $g = 0 \to 1$, as compared with **?**, will be enforced in the rest of the paper. Such a change will be shown to significantly impact the numerical experiments in Section 5.

3 Update of the ensemble

The form of the predictive prior also has important consequences on the EnKF-N theory. First of all, the pdfs Eq. (18) or Eq. (20) are multivariate T-distributions, and more specifically multivariate Cauchy distributions. They are proper, i.e. normalizable to 1, but have neither first-order nor second-order moments.

3.1 Laplace approximation

Conditioned on $\mathbf{B}$, both the prior and the posterior are Gaussian provided the observation error distribution is Gaussian which is assumed for the sake of simplicity. Without this conditioning, however, they are both a (continuous) mixture of candidate Gaussians in the EnKF-N derivation. Therefore, the posterior $p(\mathbf{w}|\mathbf{y}) \propto p(\mathbf{y}|\mathbf{w})p(\mathbf{w})$ should be interpreted with caution. As was done in **?**, its mode can in principle be safely estimated. However, its moments do not generally exist. They exist only if the likelihood $p(\mathbf{y}|\mathbf{w})$ enables it. Even when they do exist, they do not carry the same significance as for Gaussians.

Hence, the analysis $\mathbf{w}_a$ is safely defined using the EnKF-N Cauchy prior as the most likely $\mathbf{w}$ of the posterior pdf. But, using the mean and the error covariance matrix of the posterior is either impossible or questionable because as explained above they may not exist.

One candidate Gaussian that does not involve integrating over the hyperprior, is the Laplace approximation of the posterior (see **?**, for instance), which is the Gaussian approximation fitted to the pdf in the neighborhood of $\mathbf{w}_a$. This way, the covariance matrix of the Laplace distribution is obtained as the Hessian of the cost function at the minimum $\mathbf{w}_a$. Refining the covariance matrix from the inverse Hessian is not an option since the exact covariance matrix of the posterior pdf may not exist. This is a counterintuitive argument against looking for a better approximation of the posterior covariance matrix rather than the inverse Hessian.

Once a candidate Gaussian for the posterior has been obtained, the updated ensemble of the EnKF-N is obtained from the Hessian, just as in the ETKF. The updated ensemble is

$$\mathbf{E}^a = \mathbf{x}^a\mathbf{1}^\mathsf{T} + \mathbf{X}_a\,, \qquad \mathbf{x}^a = \overline{\mathbf{x}} + \mathbf{X}\mathbf{w}_a\,. \tag{22}$$

where $\mathbf{x}^a$ is the analysis in state space; $\mathbf{w}_a$ is the argument of the minimum of Eq. (21). The updated ensemble of perturbations $\mathbf{X}_a$ is given by

$$\mathbf{X}_a = \sqrt{N-1}\mathbf{X}\left[\mathcal{H}_a\right]^{-1/2}\mathbf{U}\,, \tag{23}$$

where $\mathbf{U}$ is an arbitrary orthogonal matrix satisfying $\mathbf{U1} = \mathbf{1}$ (**?**) and where $\mathcal{H}_a$ is the Hessian of Eq. (21),

$$\mathcal{H}_a = \mathbf{Y}^\mathsf{T}\mathbf{R}^{-1}\mathbf{Y} + (N+1)\frac{\left(\varepsilon_N + \mathbf{w}_a^\mathsf{T}\mathbf{w}_a\right)\mathbf{I}_N - 2\mathbf{w}_a\mathbf{w}_a^\mathsf{T}}{\left(\varepsilon_N + \mathbf{w}_a^\mathsf{T}\mathbf{w}_a\right)^2} \tag{24}$$

with $\mathbf{Y} = \mathbf{HX}$ and $\mathbf{H}$ the tangent linear of H. The algorithm of this so-called *primal* EnKF-N is recalled by Algorithm 1. Note that the algorithm can handle nonlinear observation operator since it is based on a variational analysis similarly to the maximum likelihood ensemble filter of **?**. We will choose $\mathbf{U}$ to be the identity matrix in all numerical illustrations of this paper, and in particular Section 5, in order to minimize the displacement in the analysis (**?**).

3.2 Theoretical equivalence between the primal and the dual approaches

? showed that the functional Eq. (21) is generally non-convex but has a global minimum. Yet, the cost function is only truly non-quadratic in the direction of the radial degree of freedom $\|\mathbf{w}\|$ of $\mathbf{w}$, because the predictive prior is elliptical. This remark led **?** (later **?**) to show, assuming H is linear or linearized, that the minimization of Eq. (21) can be performed simply by minimizing the following dual cost function over $]0, (N+1)/\varepsilon_N]$:

$$\mathcal{D}(\zeta) = \frac{1}{2}\boldsymbol{\delta}^\mathsf{T}\left(\mathbf{R} + \mathbf{Y}\zeta^{-1}\mathbf{Y}^\mathsf{T}\right)^{-1}\boldsymbol{\delta} + \frac{\varepsilon_N\zeta}{2} + \frac{N+1}{2}\ln\frac{N+1}{\zeta} - \frac{N+1}{2}\,, \tag{25}$$

where $\boldsymbol{\delta} = \mathbf{y} - H(\overline{\mathbf{x}})$. Its global minimum can easily be found since $\zeta \mapsto \mathcal{D}(\zeta)$ is a scalar cost function. The variable ζ is conjugate to the square radius $\|\mathbf{w}\|^2$. It can be seen as the number of effective degrees of freedom in the ensemble. Once the argument of the

minimum of $\mathcal{D}(\zeta)$, ζ_a, is computed, the analysis for $\mathbf{w}$ can be obtained from the ETKF-like cost function:

$$\mathcal{J}(\mathbf{w}) = \frac{1}{2}\left\|\mathbf{y} - H(\overline{\mathbf{x}} + \mathbf{X}\mathbf{w})\right\|^2_{\mathbf{R}} + \frac{\zeta_a}{2}\left\|\mathbf{w}\right\|^2, \tag{26}$$

with the solution:

$$\mathbf{w}_a = \left(\mathbf{Y}^\mathsf{T}\mathbf{R}^{-1}\mathbf{Y} + \zeta_a \mathbf{I}_N\right)^{-1}\mathbf{Y}^\mathsf{T}\mathbf{R}^{-1}\boldsymbol{\delta} = \mathbf{Y}^\mathsf{T}\left(\zeta_a\mathbf{R} + \mathbf{Y}\mathbf{Y}^\mathsf{T}\right)^{-1}\boldsymbol{\delta}. \tag{27}$$

Based on this effective cost function, an updated set of perturbations can be obtained:

$$\mathbf{X}_a = \sqrt{N-1}\mathbf{X}\left[\mathcal{H}_a\right]^{-\frac{1}{2}}\mathbf{U} \quad \text{with} \quad \mathcal{H}_a = \mathbf{Y}^\mathsf{T}\mathbf{R}^{-1}\mathbf{Y} + \zeta_a\mathbf{I}_N. \tag{28}$$

As a consequence, the EnKF-N with an analysis performed in ensemble space can be seen as an ETKF with an adaptive *optimal* inflation factor λ^a applied on the prior distribution, and related to ζ_a by $\lambda^a = \sqrt{(N-1)/\zeta_a}$. Provided one subscribes to the EnKF-N formalism, this tells us that sampling errors can be cured by *multiplicative* inflation. This is supported by **?** who experimentally showed that multiplicative inflation is well suited to account for sampling errors whereas additive inflation is better suited to account for model errors in a meteorological context. Other efficient adaptive inflation methods have been proposed by, e.g. **????????** for broader uses including extrinsic model error. Nevertheless, for the experiments described in Section 5, they are not as performant with the specific goal of accounting for sampling errors as the EnKF-N.

Equation (28), on which the results of **?** are based, is only an approximation because it does not use the Hessian of the complete cost function Eq. (21). Only the diagonal term of the Hessian of the background term is kept:

$$\mathcal{H}_b \simeq \frac{N+1}{\varepsilon_N + \left\|\mathbf{w}_a\right\|^2}\mathbf{I}_N, \tag{29}$$

which can be simply written $\mathcal{H}_b \simeq \zeta_a\mathbf{I}_N$ using $\zeta_a = \frac{N+1}{\varepsilon_N + \|\mathbf{w}_a\|^2}$ shown in **?** to be one of the optimum conditions. The off-diagonal rank-one correction, $-2(N+1)^{-1}\zeta_a^2\mathbf{w}_a\mathbf{w}_a^\mathsf{T}$, has been

neglected. This approximation is similar to that of the Gauss-Newton method which is an approximation of the Newton method where the Hessian of the cost function to be minimized is approximated by the product of first-order derivative terms and by neglecting second-order derivative terms. The approximation actually consists in neglecting the co-dependence of the errors in the radial ($\|\mathbf{w}\|$) and angular ($\mathbf{w}/\|\mathbf{w}\|$) degrees of freedom of $\mathbf{w}$.

Since the dual EnKF-N is meant to be equivalent to the primal EnKF-N when the observation operator is linear, the updated ensemble should actually be based on Eq. (24) which can also be written

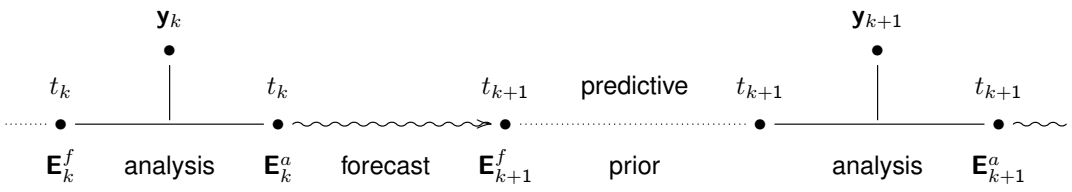
$$\mathbf{X}_a = \sqrt{N-1}\mathbf{X}\left[\mathcal{H}_a\right]^{-\frac{1}{2}}\mathbf{U} \quad \text{with} \quad \mathcal{H}_a = \mathbf{Y}^{\mathsf{T}}\mathbf{R}^{-1}\mathbf{Y} + \zeta_a\mathbf{I}_N - \frac{2\zeta_a^2}{N+1}\mathbf{w}_a\mathbf{w}_a^{\mathsf{T}}, \tag{30}$$

and compared to the approximation Eq. (28) used in **?**. The algorithm of this so-called *dual* EnKF-N is recalled in Algorithm 2 and includes the correction. With Eq. (30), the dual scheme is strictly equivalent to the primal scheme provided that H is linear, whereas it is only approximately so with Eq. (28).

The co-dependence of the radial and angular degrees of freedom exposed by the dual cost function is further explored in Appendix A.

4 Cycling of the EnKF-N and impact of model nonlinearity

We have discussed and amended the analysis step of the EnKF-N. To complete the data assimilation cycle, the ensemble must be forecasted between analyses. The cycling of the EnKF-N can be summarized by the following diagram:



In accounting for sampling error, the EnKF-N framework differs quite significantly from that of **???**. Focusing on the bias of the EnKF gain and precision matrix, these studies are geared towards single-cycle corrections. By contrast, the EnKF-N enables the likelihood to influence the estimation of the posterior covariance matrix. This can be seen by writing and recognizing the posterior as a non-uniform mixture of Gaussians, as for the prior. The inclusion of the likelihood is what makes the EnKF-N equipped to handle the effects of model nonlinearity and the sequentiality of data assimilation.

Assuming linear evolution and observation models that are taken as perfect, and provided the ensemble is big enough to span the unstable and neutral subspace, and even though it remains degenerate, inflation or localization are unnecessary in the ensemble square root Kalman filter (**??**). Sampling errors, if present, can be ignored in this case. Therefore, it is inferred from this result that inflation is actually compensating for the misestimation of errors generated by model nonlinearity. Following this line of thought, **?** hypothesized that the finite-size scheme actually accounts for the error generated in the nonlinear deformation of the ensemble in the forecast step of the EnKF.

A recent study by **?** confirms and clarifies this suggestion. The authors show that the nonlinear evolution of the error in the extended Kalman filter generates additional errors unaccounted for by the extended Kalman filter linear propagation of the error. In a specific example, they are able to avoid the need for inflation with the 40-variable Lorenz-95 model using a total of 24 perturbations (14 for the unstable and neutral subspace and 10 for the main nonlinear corrections). We checked that the same root mean square errors as shown in table II of **?** can be achieved by the EnKF-N and the optimally tuned EnKF with an ensemble of size $N = 24$. This reinforces the idea that the EnKF-N accounts, albeit within ensemble space, for the error generated by nonlinear corrections inside and outside the ensemble subspace. Additionally, note that the EnKF-N does not show any sign of divergence in the regime studied by **?** even for much stronger model nonlinearity.

To picture the impact of inflation on the fully cycled EnKF, let us consider the simplest possible, one-variable, perfect, linear model $x_{k+1} = \alpha x_k$, with k the time index. If $\alpha^2 > 1$, the model is unstable, and stable if $\alpha^2 < 1$. In terms of uncertainty quantification, multiplicative

inflation is meant to increase the errors covariances so as to account for misestimated errors. Here, we apply the inflation on the prior at each analysis step since the EnKF-N implicitly does it. Let us denote b_k the forecast/prior error variance, r the static observation error variance and a_k the error analysis variance. ζ plays the same role as in the EnKF-N scheme, so that a uniform inflation is $\zeta^{-\frac{1}{2}}$. Sequential data assimilation implies the following recursions for the variances:

$$a_k^{-1} = \zeta b_k^{-1} + r^{-1} \quad \text{and} \quad b_{k+1} = \alpha^2 a_k \,, \tag{31}$$

whose asymptotic solution ($a \equiv a_\infty$) is

$$\text{if } \alpha^2 < \zeta : a = 0 \quad \text{and} \quad \text{if } \alpha^2 \geq \zeta : a = (1 - \zeta/\alpha^2) r \,. \tag{32}$$

Now, consider a multivariate model which is the collection of several independent one-variable models with as many growth factors α. In the absence of inflation, $\zeta = 1$, the stable modes, $\alpha^2 < 1$, converge to a perfect analysis ($a = 0$) whereas the unstable modes, $\alpha^2 > 1$, converge to a finite error ($a > 0$) that grows with the instability of the modes, as expected. When inflation is used, $\zeta < 1$, the picture changes but mostly affect the modes close to neutral (see Fig. 2). The threshold is displaced and the modes with finite asymptotic errors now include a fraction of the stable modes. The strongly unstable modes are much less impacted.

In spite of its simplicity and its linearity, this model makes the link between the EnKF-N, multiplicative inflation and the dynamics. **??** have argued that, in the absence of model error, systematic error of the EnKF comes from the error transported from the unstable subspace to the stable subspace by the effect of nonlinearity. Unaccounted error would accumulate on the stable modes close to neutrality. As seen above, the use of the EnKF-N, or multiplicative inflation on the prior, precisely acts on these modes by increasing their error statistics without affecting the most unstable modes that mainly drive the performance of the EnKF.

5 Numerical experiments

Twin experiments using a perfect model and the EnKF-N have been carried out on several low-order models in previous studies. In many cases the EnKF-N, or its variant with localization (using domain localization), were reported to perform on the Lorenz-63 and Lorenz-95 models as well as the ETKF but with optimally tuned uniform inflation. With a two-dimensional forced turbulence model, driven by the barotropic vorticity advection equation, it was found to perform almost as well as the ETKF with optimally tuned uniform inflation (**?**), although the local EnKF-N has not yet been thoroughly tested with this model.

The choice of ε_N has remained a puzzle in these experiments. It has been reported that the Lorenz-63 model required $\varepsilon_N = 1 + 1/N$, whereas the Lorenz-95 model required $\varepsilon_N = 1$, seemingly owning to the larger ensemble size. It was also previously reported that domain localization of the EnKF-N with both models required $\varepsilon_N = 1 + 1/N$. In the present study, we have revisited those experiments using the correction $g = 0 \to 1$ of Section 2.4, sticking with the theoretical value $\varepsilon_N = 1 + 1/N$, and the same ensemble sizes. This essentially reproduced the results of the best choice for ε_N in each case. For these low-order models, this solved a puzzle: there is no need to adjust $\varepsilon_N = 1 + 1/N$. Hence, the EnKF-N in the subsequent experiments uses the correction $g = 0 \to 1$ and $\varepsilon_N = 1 + 1/N$.

Figure 3 summarizes the corrections of Sections 2 and 3. It also illustrates the equivalence between the primal and the dual EnKF-N. It additionally shows the performance of the dual EnKF-N with the approximate Hessian used in **?**, and the performance of the ensemble square root Kalman filter with optimally tuned uniform inflation. The Lorenz-95 low-order model is chosen for this illustration (**?**). Details about the model can be found in their article. A twin experiment is performed, with a fully observed system ($H = \mathbf{I}_d$, where $d = M = 40$), an observation error variance matrix $\mathbf{R} = \mathbf{I}_d$ which is also used to generate synthetic observations from the truth. The ensemble size is $N = 20$. The time interval between observation updates Δt is varied which changes the nonlinearity strength. Varying model nonlinear is highly relevant because, as explained in Section 4, model nonlinearity is the profound cause of the need for inflation, in this rank-sufficient context ($N = 20$). We

plot the mean analysis root mean square error (RMSE) between the analysis state and the truth state. To obtain a satisfying convergence of the statistics, the RMSEs are averaged over 10^5 cycles, after a spin-up of 5×10^3 cycles.

The performances of the primal and the dual EnKF-N are indistinguishable for the full Δt range. The dual EnKF-N with approximate Hessian hardly differs from the EnKF-N, i.e. using Eq. (28) in place of Eq. (30). However, it is slightly suboptimal for $\Delta t = 0.05$ by about 5%.

Similar experiments have been conducted with the Lorenz-63 model (**?**), the Lorenz-05II model (**?**) model, the Kuramato-Shivashinski model (**??**). These experiments have yielded the same conclusions.

The additional numerical cost of using the finite-size formalism based on Jeffreys' hyper-prior is now compared to the analysis step of an ensemble Kalman filter or of an ensemble Kalman smoother based on the ensemble-transform formulation. The computational cost depends on the type of methods. Let us first discuss non-iterative methods, such as the ETKF or a smoother based on the ETKF. If the singular value decomposition (SVD) of $\mathbf{R}^{-\frac{1}{2}}\mathbf{Y}$ has already been obtained, the dual approach can be used and the additional cost of the EnKF-N, or EnKS-N, is due to the minimization of the dual cost function Eq. (25), which is negligible. This is indeed the case in all our experiments where the SVD has been obtained in order to compute the inverse in the state update Eq. (27) or the inverse square root in the perturbations update Eq. (30) or Eq. (24). If the data assimilation is iterative (for significantly nonlinear models) such as the maximum likelihood ensemble filter (**?**) or the iterative ensemble Kalman smoother (**?**), then the primal approach of the finite-size scheme can be made to coincide with the iterative scheme. Examples of such integrated schemes are given in **??**. The additional cost is often negligible except if the number of expected iterations is small which is the case if the models are weakly nonlinear. However, in this case, the finite-size correction is also expected to be small with an effective inflation value close to 1.

Moreover, it is important to notice that the perturbations update as given by Eq. (30) can induce a significant extra numerical cost as compared to the update of an ETKF. Indeed the

SVD used to compute Eq. (27) cannot be directly used to compute Eq. (30) which might require another SVD. However, using the approximate scheme which consists in neglecting the off-diagonal term does not make that requirement. Even if the off-diagonal term is included in the Hessian, the inverse square root of the Hessian could be computed from the original SDV through a Sherman-Morisson update because the off-diagonal term is of rank one.

Let us finally mention that no significant additional storage cost is required by the scheme.

6 Performance in the prior-driven regime

The EnKF-N based on the Jeffreys' hyperprior was found to fail in the limit where the system is almost linear but remains nonlinear (**?**). This regime is rarely explored with low-order models but it is likely to be encountered in less homogeneous, more realistic applications. Figure 4a illustrates this failure. It extrapolates the results of Fig. 3 to very small time intervals between updates where the dynamics are quasi-linear. As Δt decreases the RMSE of the optimal inflation EnKF decreases as one would expect, while the RMSE of the EnKF-N based on the Jeffreys' prior increases.

In this regime, the EnKF-N has great confidence in the prior as any filter would do. Therefore, the innovation-driven term becomes less important than the prior term $\mathcal{D}_b(\zeta) = \frac{\varepsilon_N \zeta}{2} + \frac{N+1}{2} \ln \frac{N+1}{\zeta} - \frac{N+1}{2}$ in the dual cost function Eq. (25), so that its mode ζ_a tends to the mode of $\mathcal{D}_b(\zeta)$ which is $\zeta_a = (N+1)/\varepsilon_N = N$. Note that an inflation of 1 corresponds to $\zeta = N - 1$. Hence, in this regime, even for moderately-sized innovations, there is deflation. The failure of the EnKF-N was empirically fixed in **?** by capping ζ_a to prevent deflation.

More generally, we believe the problem is to be encountered whenever the prior largely dominates the analysis (prior-driven regime). This is bound to happen when the observations are too few, too sparsely distributed, which could occur when using domain localization, and whenever they are unreliable compared to the prior. Quasi-linear dynamics also fit this description, the ratio of the observation precision to the prior precision becoming small after a few iterations.

This failure may not be due to the EnKF-N framework. It may be due to an inappropriate choice of candidate Gaussian posterior as described in Sec. 3. Or it may be due to an inappropriate choice of hyperprior in this regime. Although it seems difficult to devise a hyperprior that performs optimally in all regimes, we can suggest two adjustments to Jeffreys' hyperprior in this prior-driven regime.

6.1 Capping of the inflation

Here, deflation is avoided by capping ζ. Firstly, we build the desired dual cost function. Instead of minimizing $\mathcal{D}(\zeta)$ over $]0,(N+1)/\varepsilon_N]$, it is minimized over $]0,\overline{\zeta}]$, with $0 \leq \overline{\zeta} \leq (N+1)/\varepsilon_N$, which defines the dual cost function. $\overline{\zeta}$ is a tunable bound which is meant to be fixed over a wide range of regimes. Following a similar derivation to Appendix A of **?**, one can show that the background term of the primal cost function corresponding to this dual cost function is

$$\begin{aligned}
&\text{if } \|\mathbf{w}\|^2 \leq \frac{N+1}{\overline{\zeta}} - \varepsilon_N : \quad \mathcal{J}_b(\mathbf{w}) = \frac{\overline{\zeta}}{2}\left(\varepsilon_N + \|\mathbf{w}\|^2\right) + \frac{N+1}{2}\ln\left(\frac{N+1}{\overline{\zeta}}\right) - \frac{N+1}{2} \\
&\text{if } \|\mathbf{w}\|^2 > \frac{N+1}{\overline{\zeta}} - \varepsilon_N : \quad \mathcal{J}_b(\mathbf{w}) = \frac{N+1}{2}\ln\left(\varepsilon_N + \|\mathbf{w}\|^2\right) .
\end{aligned} \tag{33}$$

The dual and primal cost functions can both be shown to be convex. There is no duality gap, which means, with our definitions of these functions, that the minimum of the dual cost function is equal to the minimum of the primal cost function. By construction, in the small innovation range, i.e. $\|\mathbf{w}\|^2 \leq (N+1)/\overline{\zeta} - \varepsilon_N$, the EnKF-N, endowed with this new hyperprior, corresponds to the ETKF (**?**) with an inflation of the prior by $(N-1)/\overline{\zeta} \geq 1$. Since the hyperprior assumed in the regime of small $\|\mathbf{w}\|$ is $p(\mathbf{x}_b, \mathbf{B}) = \delta(\mathbf{B} - \overline{\zeta}\mathbf{P})$, this could be called the Dirac-Jeffreys hyperprior.

Even with the Dirac-Jeffreys hyperprior, it is still necessary to introduce a tiny amount of inflation through $\overline{\zeta}$ in the quasi-linear regime. This might prove barely relevant in a high-dimensional realistic system as it was for the sensitive low-order models that we tested the scheme with. Even with Lorenz-95, an instability develops over very long experimental runs

in the absence of this residual inflation. Still this remains a theoretical concern. Moreover, we could not find a rigorous argument to support avoiding deflation in all regimes, and hence the capping. That is why we propose an alternative solution in the following.

6.2 Smoother schemes in the prior-driven regime

In the limit of **R** getting very large, the observations cannot carry information, and the ensemble should not be updated at all, i.e. it should be close to the prior ensemble, with an inflation of 1 ($\zeta = N-1$). Outside of this regime, we do not see any fundamental reason to constrain ζ to be smaller than $N-1$. A criterion to characterize this regime would be

$$\psi = \frac{1}{N-1}\mathsf{Tr}\left(\mathbf{Y}^\mathsf{T}\mathbf{R}^{-1}\mathbf{Y}\right), \tag{34}$$

which computes the ratio of the prior variances to the observation error variances. When ψ tends to zero, the analysis should be dominated by the prior and ζ should tend to $N-1$. When ψ drifts away from zero, we do not want to alter the hyperprior and the EnKF-N scheme, even if it implies deflation. We found several schemes that satisfy these constraints. Two of them, denoted R1 and R2, consist in modifying ε_N into ε'_N and yield a well-behaved mode of the background part of the dual cost function $\zeta_b = \underset{\zeta}{\text{argmin}}\,[\mathcal{D}_b(\zeta)]$:

$$\begin{aligned}
&\text{R1:}\quad \varepsilon'_N = \frac{\varepsilon_N}{1-\frac{1}{N}e^{-\psi}} && \implies \quad \zeta_b = N - e^{-\psi} \\
&\text{R2:}\quad \varepsilon'_N = \frac{N+1}{N}\left(\frac{N}{N-1}\right)^{\frac{1}{1+\psi}} && \implies \quad \zeta_b = N\left(\frac{N-1}{N}\right)^{\frac{1}{1+\psi}}
\end{aligned} \tag{35}$$

The point of these formulae is to make ζ_b tend to $N-1$ (no inflation) when the criterion ψ tends to zero. On the other hand, when ψ gets bigger ζ_b tends to N, i.e. to the original dual cost function's behavior dictated by Jeffreys' hyperprior. The implementation of these schemes is straightforward for any of the Algorithms 1 or 2, since only ε_N needs to be modified either in the dual or the primal cost functions.

6.3 Numerical illustrations

The performance of the Dirac-Jeffreys EnKF-N where we choose $\sqrt{(N-1)/\overline{\zeta}} = 1.005$, and of the EnKF-N with the hyperprior corrections (R1) and (R2), are illustrated with a twin experiment on the Lorenz-95 model in the quasi-linear regime. Also included are the EnKF-N with Jeffreys' prior and the ensemble square root Kalman filter with optimally tuned inflation. The RMSEs are plotted as a function of Δt in $[0.01, 0.5]$ in Fig. 4a.

Another way to make a data assimilation system based on the Lorenz-95 more linear, rather than decreasing Δt, is to decrease the forcing parameter to render the model more linear. Figure 4b illustrates this when F is varied from 4 (linear) to 12 (strongly nonlinear), with $\Delta t = 0.05$, and the same set-up as in Section 5. As anticipated, the EnKF-N based on Jeffreys' hyperprior fails for $F < 7.5$. However, the EnKF-N based on the Dirac-Jeffreys' hyperprior and the EnKF-N with the schemes R1 and R2 show performances equivalent to the EnKF with optimally tuned inflation. We remark a slight underperformance of the EnKF-N in the very strongly chaotic regimes compared to the optimally tuned EnKF. We have also check that these good performances also apply to the Lorenz-63 model.

The spread of the ensemble for the Dirac-Jeffreys EnKF-N has also been plotted in Fig. 4a and Fig. 4b. The value of the spread is consistent with the RMSE except in significantly nonlinear regimes such as when $\Delta t > 0.15$ and $F = 8$, or to a lesser extent when $\Delta t = 0.05$ and $F > 8$. In those nonlinear regimes and with such non-iterative EnKFs, the Gaussian error statistics approximation is invalidated so that the RMSE could differ significantly from the ensemble spread.

7 Informative hyperprior, covariance localization and hybridization

So far, the EnKF-N has relied on a noninformative hyperprior. In this section we examine, mostly at a formal level, the possibility to account for additional, possibly independent, information on the error statistics, like an hybrid EnKF-3D-Var is meant to (**??**). A single

numerical illustration is intended since extended results would involve much more developments and would be very model-dependent.

In a perfect model context, we observed that uncertainty on the variances usually addressed by inflation could be taken care of by the EnKF-N based on Jeffreys' hyperprior. However, it does not take care of the correlation (as opposed to variance) and rank-deficiency issues, which are usually addressed by localization. Localization has to be superimposed to the finite-size scheme to build a local EnKF-N without the intrinsic need for inflation (**?**). Nonetheless, by marginalizing over limited-range covariance matrices (Section 5 of **?**), we also argued that the use of an informative hyperprior would produce covariance localization within the EnKF-N framework. A minimal example where the hyperprior is defined over $\mathbf{B}$ matrices that are positive diagonal, hence very short-ranged, was given and supported by a numerical experiment. Hence, it is likely that the inclusion of informative prior is a way to elegantly impose localization within the EnKF-N framework.

An informative hyperprior is the normal-inverse-Wishart (NIW) pdf:

$$p_{\text{NIW}}(\mathbf{x}_b, \mathbf{B}) \propto |\mathbf{B}|^{-\frac{M+2+\nu}{2}} \exp\left[-\frac{\kappa}{2}\left\|\mathbf{x}_b - \mathbf{x}_c\right\|^2_{\mathbf{B}} - \frac{1}{2}\text{Tr}\left(\mathbf{B}^{-1}\mathbf{C}\right)\right]. \tag{36}$$

It is convenient because, with this hyperprior, Eq. (3) remains analytically integrable. The location state $\mathbf{x}_c$, the scale matrix $\mathbf{C}$, which is assumed to be full-rank, κ and ν are hyperparameters of the distribution from which the true error moments $\mathbf{x}_b$ and $\mathbf{B}$ are drawn. The pdf p_{NIW} is proper only if $\nu > M-1$, but this is not an imperative requirement provided that the integral in Eq. (3) is proper.

The resulting predictive prior can be deduced from **?** Section 3.6:

$$p(\mathbf{x}|\mathbf{E}) \propto \left\{1 + \frac{N+\kappa}{N+\kappa+1}\left\|\mathbf{x} - \hat{\mathbf{x}}\right\|^2_{\frac{\kappa N}{N+\kappa}(\mathbf{x}_c-\overline{\mathbf{x}})(\mathbf{x}_c-\overline{\mathbf{x}})^{\mathsf{T}}+\mathbf{X}\mathbf{X}^{\mathsf{T}}+\mathbf{C}}\right\}^{-\frac{1}{2}(N+1+\nu)} \tag{37}$$

where $\hat{\mathbf{x}} = (\kappa\mathbf{x}_c + N\overline{\mathbf{x}})/(N+\kappa)$. From these expressions, $\mathbf{x}_c$ could be interpreted as some climatological state and $\mathbf{C}$ would be proportional to some error covariance matrix, which could be estimated from a prior, long and well-tuned EnKF run. They could also be pa-

rameterized by tunable scalars that could be estimated by a maximum likelihood principle (**?**).

A subclass of hyperpriors is obtained when the degree of freedom $\mathbf{x}_c$ is taken out, leading to the inverse Wishart (IW) distribution:

$$p_{\mathrm{IW}}(\mathbf{x}_b, \mathbf{B}) \quad \propto |\mathbf{B}|^{-\frac{M+1+\nu}{2}} \exp\left[-\tfrac{1}{2}\mathrm{Tr}\left(\mathbf{B}^{-1}\mathbf{C}\right)\right], \tag{38}$$

and to the predictive prior

$$p(\mathbf{x}|\mathbf{E}) \propto \left\{1 + \tfrac{N}{N+1}\,\|\mathbf{x}-\overline{\mathbf{x}}\|^2_{\mathbf{X}\mathbf{X}^{\mathsf{T}}+\mathbf{C}}\right\}^{-\frac{1}{2}(N+\nu)}. \tag{39}$$

Jeffreys' hyperprior is recovered from the IW hyperprior in the limit where $\nu \to 0$ and $\mathbf{C} \to \mathbf{0}$, well within the region $\nu \leq M-1$ where the IW pdf is improper. Note that the use of an IW distribution was advocated owing to its natural conjugacy in a remarkable paper by **?** where a hierarchical stochastic EnKF was first proposed and developed.

Because the scale matrix $\mathbf{C}$ is assumed full-rank, updating in state space is preferred to an analysis in ensemble space. Based on the marginals Eq. (37) and Eq. (39), the $\mathcal{J}_b$ term of the analysis cost function is of the form:

$$\mathcal{J}_b(\mathbf{x}) = \frac{\gamma}{2}\ln\left[\varepsilon_N + \|\mathbf{x}-\hat{\mathbf{x}}\|^2_{\mathbf{\Gamma}}\right] \quad \text{with} \quad \mathbf{\Gamma} = \mathbf{X}\mathbf{X}^{\mathsf{T}} + \widehat{\mathbf{C}}. \tag{40}$$

In the case of the NIW hyperprior, one has:

$$\gamma = N+1+\nu, \qquad \varepsilon_N = 1+1/(N+\kappa), \qquad \widehat{\mathbf{C}} = \mathbf{C} + \frac{\kappa N}{N+\kappa}\left(\mathbf{x}_c-\overline{\mathbf{x}}\right)\left(\mathbf{x}_c-\overline{\mathbf{x}}\right)^{\mathsf{T}}. \tag{41}$$

In the case of the IW hyperprior, one has:

$$\gamma = N+\nu, \qquad \varepsilon_N = 1+1/N, \qquad \hat{\mathbf{x}} = \overline{\mathbf{x}}, \qquad \widehat{\mathbf{C}} = \mathbf{C}. \tag{42}$$

We observe that the $\mathcal{J}_b$ term is formally similar to that of the EnKF-N with Jeffreys' hyperprior which is directly obtained in state space from Eq. (7). Hence the sequential data assimilation schemes built from the NIW and IW hyperpriors formally follow that of the EnKF-N. But, to do so, the analysis must be written in state space, whereas it has been expressed in ensemble space so far.

7.1 Primal analysis and dual analysis

The primal analysis in state space is obtained from $\mathbf{x}_a = \text{argmin}_{\mathbf{x}} \mathcal{J}(\mathbf{x})$, where

$$\mathcal{J}(\mathbf{x}) = \mathcal{J}_o(\mathbf{x}) + \mathcal{J}_b(\mathbf{x}) = \frac{1}{2} \|\mathbf{y} - H(\mathbf{x})\|^2_{\mathbf{R}} + \frac{\gamma}{2} \ln \left[\varepsilon_N + \|\mathbf{x} - \hat{\mathbf{x}}\|^2_{\mathbf{\Gamma}} \right] . \tag{43}$$

For the dual analysis, we further assume that the observation operator **H** is linear (for the primal/dual correspondence to be exact). The derivation of the dual cost function follows that of **?**. The following Lagrangian is introduced to separate the radial and angular degrees of freedom of **x**:

$$\mathcal{L}(\mathbf{x}, \rho, \zeta) = \mathcal{J}_o(\mathbf{x}) + \frac{\zeta}{2} \left[\|\mathbf{x} - \hat{\mathbf{x}}\|^2_{\mathbf{\Gamma}} - \rho \right] + \frac{\gamma}{2} \ln \left(\varepsilon_N + \rho \right) . \tag{44}$$

where ζ is a Lagrange multiplier. The saddle-point equations of this Lagrangian are:

$$\rho^a = \|\mathbf{x}_a - \hat{\mathbf{x}}\|^2_{\mathbf{\Gamma}} , \tag{45}$$

$$\rho^a = \frac{\gamma}{\zeta_a} - \varepsilon_N , \tag{46}$$

$$\mathbf{x}_a = \hat{\mathbf{x}} + \mathbf{\Gamma H}^{\mathsf{T}} \left(\zeta_a \mathbf{R} + \mathbf{H \Gamma H}^{\mathsf{T}} \right)^{-1} \hat{\boldsymbol{\delta}} \quad \text{with} \quad \hat{\boldsymbol{\delta}} = \mathbf{y} - \mathbf{H}\hat{\mathbf{x}} . \tag{47}$$

$\mathbf{x}_a$, ρ^a, and ζ_a are the saddle-point values of the variables. Using these saddle-point equations, it can be shown that the minimization of Eq. (43) is equivalent to the minimization of the following scalar dual cost function over $]0, \gamma/\varepsilon_N]$

$$\mathcal{D}(\zeta) = \mathcal{L}(\mathbf{x}_a, \rho^a, \zeta) = \frac{1}{2} \hat{\boldsymbol{\delta}}^{\mathsf{T}} \left(\mathbf{R} + \zeta^{-1} \mathbf{H \Gamma H}^{\mathsf{T}} \right)^{-1} \hat{\boldsymbol{\delta}} + \frac{\varepsilon_N \zeta}{2} + \frac{\gamma}{2} \ln \frac{\gamma}{\zeta} - \frac{\gamma}{2} , \tag{48}$$

a mild generalization of Eq. (25). As in **?**, ζ is interpreted as an effective size of the ensemble as seen by the analysis. Note that, in this context, it could easily be larger than $N - 1$ if the added information content of the informative hyperprior is significant.

7.2 State space update of the ensemble perturbations

Recall that the square root ensemble update corresponding to Eq. (30) and Jeffreys' hyperprior is

$$\mathbf{X}_a = \sqrt{N-1}\mathbf{X}\left[\mathbf{Y}^\mathsf{T}\mathbf{R}^{-1}\mathbf{Y} + \zeta_a \mathbf{I}_N - \frac{2\zeta_a^2}{N+1}\mathbf{w}_a\mathbf{w}_a^\mathsf{T}\right]^{-\frac{1}{2}}\mathbf{U}\,. \tag{49}$$

Note that covariance localization cannot be implemented in ensemble space using Eq. (49). To make the covariance matrix explicit, we wish to write this in state space. Firstly, from Eq. (27), $\mathbf{w}_a$ can be written $\mathbf{w}_a = \mathbf{Y}^\mathsf{T}\mathbf{z}$, where $\mathbf{z} = \left(\zeta_a\mathbf{R} + \mathbf{Y}\mathbf{Y}^\mathsf{T}\right)^{-1}\boldsymbol{\delta}$. Then, by the matrix shift lemma which asserts that $\mathbf{A}f(\mathbf{B}\mathbf{A}) = f(\mathbf{A}\mathbf{B})\mathbf{A}$ for any two matrices $\mathbf{A}$ and $\mathbf{B}$ of compatible sizes and f an analytic function[1], we can turn this right-transform into a left-transform[2]:

$$\mathbf{X}_a = \sqrt{N-1}\left[\zeta_a\mathbf{I}_M + \mathbf{X}\mathbf{Y}^\mathsf{T}\left(\mathbf{R}^{-1} - \frac{2\zeta_a^2}{N+1}\mathbf{z}\mathbf{z}^\mathsf{T}\right)\mathbf{H}\right]^{-\frac{1}{2}}\mathbf{X}\mathbf{U}\,. \tag{50}$$

When $\zeta_a = N-1$ and $\mathbf{z} = \mathbf{0}$, one recovers the ensemble square root Kalman update formula written in state space: $\mathbf{X}_a = \left[\mathbf{I}_M + \mathbf{P}\mathbf{H}^\mathsf{T}\mathbf{R}^{-1}\mathbf{H}\right]^{-\frac{1}{2}}\mathbf{X}$ (**?**). Note that we could absorb $-\frac{2\zeta_a^2}{N+1}\mathbf{z}\mathbf{z}^\mathsf{T}$ into $\mathbf{R}$ using the Sherman-Morrison formula, leading to an effective observation error covariance matrix $\mathbf{R}_e$ which is bigger than $\mathbf{R}$ (using the order of the positive symmetric matrices). To superimpose localization on this Jeffreys' hyperprior EnKF-N, a Schur product can easily be applied to $\mathbf{X}\mathbf{Y}^\mathsf{T}$ in Eq. (50), while the transformation still applies to the initial perturbations $\mathbf{X}$ without any explicit truncation.

[1] Assuming $f(x) = \sum_{k=0}^{\infty} a_k x^k$, one has $\mathbf{A}f(\mathbf{B}\mathbf{A}) = \sum_{k=0}^{\infty} a_k \mathbf{A}(\mathbf{B}\mathbf{A})^k = \sum_{k=0}^{\infty} a_k (\mathbf{A}\mathbf{B})^k\mathbf{B} = f(\mathbf{A}\mathbf{B})\mathbf{A}$.

[2] Let $\mathbf{A}$ be a diagonalizable, non necessarily symmetric, matrix $\mathbf{A} = \boldsymbol{\Omega}\boldsymbol{\Lambda}\boldsymbol{\Omega}^{-1}$ with $\boldsymbol{\Lambda}$ diagonal. If $\boldsymbol{\Lambda} \geq \mathbf{0}$, then the square root matrix $\mathbf{A}^{\frac{1}{2}}$ is defined by $\boldsymbol{\Omega}\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{\Omega}^{-1}$.

Here, however, we wish to obtain a similar left-transform but for the NIW EnKF-N. The Hessian of the primal cost function Eq. (43) is:

$$\mathcal{H} = \mathbf{H}^{\mathsf{T}}\mathbf{R}\mathbf{H} + \gamma\frac{\mathbf{\Gamma}^{-1}}{\varepsilon_N + \|\mathbf{x}-\hat{\mathbf{x}}\|^2_{\mathbf{\Gamma}}} - 2\gamma\frac{\mathbf{\Gamma}^{-1}(\mathbf{x}-\hat{\mathbf{x}})(\mathbf{x}-\hat{\mathbf{x}})^{\mathsf{T}}\mathbf{\Gamma}^{-1}}{\left[\varepsilon_N + \|\mathbf{x}-\hat{\mathbf{x}}\|^2_{\mathbf{\Gamma}}\right]^2}, \tag{51}$$

yielding at the minimum:

$$\mathcal{H}_a = \mathbf{H}^{\mathsf{T}}\mathbf{R}^{-1}\mathbf{H} + \zeta_a\mathbf{\Gamma}^{-1} - 2\frac{\zeta_a^2}{\gamma}\mathbf{\Gamma}^{-1}(\mathbf{x}_a-\hat{\mathbf{x}})(\mathbf{x}_a-\hat{\mathbf{x}})^{\mathsf{T}}\mathbf{\Gamma}^{-1} \equiv \mathbf{H}^{\mathsf{T}}\mathbf{R}^{-1}\mathbf{H} + \zeta_a\mathbf{\Gamma}_e^{-1}, \tag{52}$$

where the correction term has been absorbed into an effective symmetric positive definite matrix $\mathbf{\Gamma}_e$. Henceforth, $\mathbf{\Gamma}$ will stand for $\mathbf{\Gamma}_e$, and any correction term is assumed to have been absorbed into $\widehat{\mathbf{C}}$ in $\mathbf{\Gamma}$. Decomposing $\zeta_a^{-1}\mathbf{\Gamma}$, which is the effective background error covariance matrix, into as many modes as required $\zeta_a^{-1}\mathbf{\Gamma} = \mathbf{Z}\mathbf{Z}^{\mathsf{T}}$ and applying Eq. (50), it is not difficult to obtain a square root matrix of the analysis error covariance matrix $\mathbf{P}_a$:

$$\mathbf{P}_a^{\frac{1}{2}} = \left[\zeta_a\mathbf{I}_M + \mathbf{\Gamma}\mathbf{H}^{\mathsf{T}}\mathbf{R}^{-1}\mathbf{H}\right]^{-\frac{1}{2}}\mathbf{\Gamma}^{\frac{1}{2}}. \tag{53}$$

However, this does not constitute a limited-size ensemble of perturbations since $\mathbf{P}_a^{\frac{1}{2}}$ is full-rank as $\mathbf{C}$ was assumed full-rank. To obtain an ensemble update of N perturbations, the scale matrix $\widehat{\mathbf{C}}$ in $\mathbf{\Gamma} = \mathbf{X}\mathbf{X}^{\mathsf{T}} + \widehat{\mathbf{C}}$ can be projected onto the ensemble space generated by the initial perturbations. Then, $\Pi_{\mathbf{X}}\widehat{\mathbf{C}}\Pi_{\mathbf{X}}$ replaces $\widehat{\mathbf{C}}$, where $\Pi_{\mathbf{X}}$ is the orthogonal projector on the columns of $\mathbf{X}$, $\Pi_{\mathbf{X}} = \mathbf{X}\mathbf{X}^{\dagger}$. Following **?**, we can form an effective set of perturbations $\mathbf{X}_c$ that satisfy

$$\mathbf{X}_c\mathbf{X}_c^{\mathsf{T}} = \mathbf{X}\mathbf{X}^{\mathsf{T}} + \Pi_{\mathbf{X}}\widehat{\mathbf{C}}\Pi_{\mathbf{X}} = \mathbf{X}\left[\mathbf{I}_N + \mathbf{X}^{\dagger}\widehat{\mathbf{C}}\left(\mathbf{X}^{\mathsf{T}}\right)^{\dagger}\right]\mathbf{X}^{\mathsf{T}} \tag{54}$$

by using

$$\mathbf{X}_c = \mathbf{X}\left[\mathbf{I}_N + \mathbf{X}^{\dagger}\widehat{\mathbf{C}}\left(\mathbf{X}^{\mathsf{T}}\right)^{\dagger}\right]^{\frac{1}{2}} \tag{55}$$

or alternatively a left-transform equivalent formula which is obtained from the matrix shift lemma

$$\mathbf{X}_c = \left[\mathbf{I}_M + \mathbf{X}\mathbf{X}^\dagger \widehat{\mathbf{C}} \left(\mathbf{X}\mathbf{X}^\mathsf{T}\right)^\dagger\right]^{\frac{1}{2}} \mathbf{X} = \left[\mathbf{I}_M + \Pi_{\mathbf{X}} \widehat{\mathbf{C}} \Pi_{\mathbf{X}} \left(\mathbf{X}\mathbf{X}^\mathsf{T}\right)^\dagger\right]^{\frac{1}{2}} \mathbf{X}. \tag{56}$$

Substituting this $\mathbf{X}_c$ to $\mathbf{\Gamma}^{\frac{1}{2}}$ in Eq. (53), we finally obtain an update of the perturbations $\mathbf{X}$ as a new set of perturbations of the same size N:

$$\mathbf{X}_a = \sqrt{N-1}\left[\zeta_a \mathbf{I}_M + \mathbf{\Gamma}\mathbf{H}^\mathsf{T}\mathbf{R}^{-1}\mathbf{H}\right]^{-\frac{1}{2}} \left[\mathbf{I}_M + \mathbf{X}\mathbf{X}^\dagger \widehat{\mathbf{C}} \left(\mathbf{X}\mathbf{X}^\mathsf{T}\right)^\dagger\right]^{\frac{1}{2}} \mathbf{X}\mathbf{U}. \tag{57}$$

7.3 Covariance localization and EnKF-3D-Var hybridization

The state space formulation of the analysis enables covariance localization which was not possible in ensemble space. To regularize $\mathbf{P} = \mathbf{X}\mathbf{X}^\mathsf{T}/(N-1)$ by covariance localization, one can apply a Schur product with a short-range correlation matrix $\mathbf{\Theta}$. In that case, Eq. (43) is unchanged but with $\mathbf{\Gamma} = \widehat{\mathbf{C}} + \mathbf{\Theta} \circ \left(\mathbf{X}\mathbf{X}^\mathsf{T}\right)$, with $\circ$ the Schur product symbol. Note that this type of covariance localization is not induced by the hyperprior, but superimposed to the EnKF-N whatever its hyperprior. The state update is obtained from Eq. (47) and Eq. (48) by letting $\mathbf{H}\mathbf{\Gamma}\mathbf{H}^\mathsf{T} \longrightarrow \mathbf{\Theta} \circ \left(\mathbf{Y}\mathbf{Y}^\mathsf{T}\right) + \mathbf{H}\widehat{\mathbf{C}}\mathbf{H}^\mathsf{T}$, or $\mathbf{\Gamma}\mathbf{H}^\mathsf{T} \longrightarrow \mathbf{\Theta} \circ \left(\mathbf{X}\mathbf{Y}^\mathsf{T}\right) + \widehat{\mathbf{C}}\mathbf{H}^\mathsf{T}$.

An alternative is to use the $\boldsymbol{\alpha}$ control variables (**??**). A mathematically equivalent cost function to Eq. (43) but with $\mathbf{\Gamma} = \widehat{\mathbf{C}} + \mathbf{\Theta} \circ \left(\mathbf{X}\mathbf{X}^\mathsf{T}\right)$ is

$$\mathcal{J}(\delta\mathbf{x}, \{\boldsymbol{\alpha}_n\}) = \mathcal{J}_o\left(\hat{\mathbf{x}} + \delta\mathbf{x} + \sum_{n=1}^{N} \boldsymbol{\alpha}_n \circ \{\mathbf{x}_n - \overline{\mathbf{x}}\}\right) + \frac{\gamma}{2} \ln\left(\varepsilon_N + \|\delta\mathbf{x}\|^2_{\widehat{\mathbf{C}}} + \sum_{n=1}^{N} \|\boldsymbol{\alpha}_n\|^2_{\mathbf{\Theta}}\right). \tag{58}$$

The $\{\boldsymbol{\alpha}_n\}_{n=1,\ldots,N}$ are N ancillary control vectors of size M related to the dynamical errors, whereas $\delta\mathbf{x}$ is a control vector of size M related to the background errors. The control vector $\mathbf{x}$ is related to $\{\boldsymbol{\alpha}_n\}$ and $\delta\mathbf{x}$ by identifying $\mathbf{x}$ with the argument of $\mathcal{J}_o$ in the cost function. This expression of the cost function is obtained by first passing from Eq. (43) to Eq. (44), then

along the lines of **?**. It can be seen from the cost function that the EnKF-N based on the NIW hyperprior yields a generalization of the EnKF-3D-Var hybrid data assimilation method to the EnKF-N framework.

Moreover, the above derivation suggests the following perturbation update needed to complete the NIW EnKF-N scheme:

$$\mathbf{X}_a = \sqrt{N-1}\left[\zeta_a \mathbf{I}_M + \left\{\widehat{\mathbf{C}}\mathbf{H}^{\mathsf{T}} + \boldsymbol{\Theta} \circ \left(\mathbf{X}\mathbf{Y}^{\mathsf{T}}\right)\right\}\mathbf{R}^{-1}\mathbf{H}\right]^{-\frac{1}{2}}\left[\mathbf{I}_M + \widehat{\mathbf{C}}\boldsymbol{\Theta} \circ \left(\mathbf{X}\mathbf{X}^{\mathsf{T}}\right)^{\dagger}\right]^{\frac{1}{2}}\mathbf{X}\mathbf{U}. \tag{59}$$

7.4 Numerical illustration

Here we wish to illustrate the use of the EnKF-N based on the IW hyperprior. We consider again the same numerical setup as in Section 5 with the Lorenz-95 model. The ν hyperparameter and the $\mathbf{C}$ scale matrix are chosen to be:

$$\nu = 1 + N\frac{\alpha}{1-\alpha}, \qquad \mathbf{C} = \frac{\beta}{1-\beta}\mathbf{I}_M \tag{60}$$

with α and β two real parameters in the interval $[0,1[$. Synthetic experiments are performed for a wide range of (α,β) couples for two sizes of the ensemble: $N = 20$, which is bigger than the dimension of the unstable and neutral subspace (14) which, for traditional EnKFs, would not require localization but inflation, and $N = 10$ which, for traditional EnKFs, would require both localization and inflation. We do not use inflation since it is meant to be accounted for by the finite-size scheme. We do not superimpose domain or covariance localization. Analysis RMSEs are computed for each run and reported in Fig. 5.

This is a preliminary experiment. In particular we do not perform any optimization of α and β based for instance on empirical Bayesian estimation. For $N = 20$, we barely remark any improvement in term of RMSEs due to the use of the NIW hyperprior as compared to the EnKF-N based on Jeffreys' hyperprior, i.e. $(\alpha,\beta) = (0,0)$. However, we observe that for $N = 10$ localization is naturally enforced via the hyperprior due to a mechanism known in statistics as *shrinkage*. Although there is no dynamical tuning of α and β, and even though the choice for $\mathbf{C}$ is gross, good RMSEs can be obtained. A RMSE of 0.33 is achieved for

$(\alpha, \beta) = (0.50, 057)$ as compared to a typical analysis RMSE of 0.20 for the EnKF-N with optimally tuned, superimposed localization. Interestingly, the average optimal effective size in this case is $\zeta_a = 15$, above the unstable subspace dimension, validating its potential use as a diagnostic.

8 Conclusions

In this article, we have revisited the finite-size ensemble Kalman filter, or EnKF-N. The scheme offers a Bayesian hierarchical framework to account for the uncertainty in the forecast error covariance matrix of the EnKF which is inferred from a limited-size ensemble. We have discussed, introduced additional arguments for, and sometimes improved several of the key steps of the EnKF-N derivation. Our main findings are:

1. A proper account of the gauge degrees of freedom in the redundant ensemble of perturbations and the resulting analysis led to a small but important modification of the ensemble transform-based EnKF-N analysis cost function ($g = 0 \to 1$, as seen in Eq. (21)).

2. Consequently, the marginal posterior distribution of the system state is a Cauchy distribution, which is proper but does not have first and second-order moments. Hence, only the maximum a posteriori estimator is unambiguously defined. Moreover, this suggests that the Laplace approximation should be used to estimate the full posterior.

3. The modification $g = 0 \to 1$ frees us from the inconvenient tweaking of ε_N to 1 or to $1 + \frac{1}{N}$: now, only $\varepsilon_N = 1 + \frac{1}{N}$ is required.

4. The connection to dynamics has been clarified. It had already been assumed that the EnKF-N compensates for the nonlinear deformation of the ensemble in the forecast step. This conjecture was here substantiated by arguing that the effect of the nonlinearities is similar to sampling error, thus explaining why multiplicative inflation, and the EnKF-N in particular, can compensate for it.

5. The ensemble update of the dual EnKF-N was amended to offer a perfect equivalence with the primal EnKF-N. It was shown that the additional term in the posterior error covariance matrix accounts for the error co-dependence between the angular and the radial degrees of freedom. However, this correction barely affected the numerical experiments we tested it with.

6. The EnKF-N based on Jeffreys' hyperprior led to unsatisfying performance in the limit where the analysis is largely driven by the prior, especially in the regime where the model is almost (but not) linear. We proposed two new types of schemes which rectify the hyperprior. These schemes have been successfully tested on low-order models, meaning that the performance of the EnKF-N becomes as good as the ensemble square root Kalman filter with optimally tuned inflation in all the tested dynamical regimes.

7. As originally mentioned in **?**, the EnKF-N offers a broad framework to craft variants of the EnKF with alternative hyperpriors. Inflation was shown to be addressed by a noninformative hyperprior whereas a localization seems to require an informative hyperprior. Here, we showed that choosing the informative normal-inverse-Wishart distribution as a hyperprior for $\mathbf{x}_b, \mathbf{B}$ leads to a formally similar EnKF-N, albeit expressed in state space rather than ensemble space. The EnKF-N based on this informative hyperprior is a finite-size variant of the hybrid EnKF-3D-Var. It has a potential for tuning the balance between static and dynamical errors. Moreover, we showed on a preliminary numerical experiment that localization can be naturally carried out through shrinkage induced by the scale matrix of the normal-inverse-Wishart hyperprior.

With the corrections and new interpretations on the EnKF-N based on Jeffreys' hyperprior, we have obtained a practical and robust tool that can be used in perfect model EnKF experiments in a wide range of conditions without the burden of tuning the multiplicative inflation. This has saved us a lot of computational time in recent published methodological studies.

An EnKF-N based on an informative hyperprior, the normal-inverse-Wishart distribution, has been described and its equations derived. We plan to evaluate it thoroughly on extensive numerical experiments. Several optional uses of the method are contemplated. Hyperparameters $\mathbf{x}_c$, $\mathbf{C}$, ν and κ could be diagnosed from the statistics of a prior well-tuned data assimilation run. Empirical Bayesian approaches could then be used to objectively balance the static errors and the dynamical errors. Alternatively, the hyperparameters could be estimated online in the course of the EnKF, rather than being obtained from prior statistics, using a more systematic empirical Bayesian approach.

Acknowledgements. This study is a contribution to the INSU/LEFE project DAVE.

Appendix A: Coupling of the radial and angular degrees of freedom

Section 3.2 separately identified angular and radial degrees of freedom in the EnKF-N cost function. This led to the dual cost function, and an alternative interpretation of the EnKF-N as an adaptive inflation scheme that accounts for sampling errors.

Here we wish to interpret the contributions in the Hessian Eq. (24) that come from the angular and from the radial degrees of freedom. To do so, we study the evidence $p(\mathbf{y})$, i.e. the likelihood of the observation vector, as estimated from the EnKF-N. This evidence is usually computed by marginalizing over all possible model states, which reads in our case:

$$p(\mathbf{y}) = \int_{\mathbb{R}^N} \mathrm{d}\mathbf{w}\, p(\mathbf{y}|\mathbf{w}) p(\mathbf{w}) = \mathcal{A}_N \int_{\mathbb{R}^N} \mathrm{d}\mathbf{w}\, e^{-\frac{1}{2}\|\mathbf{y}-H(\bar{\mathbf{x}}+\mathbf{X}\mathbf{w})\|^2_{\mathbf{R}} - \frac{N+1}{2}\ln\left(\varepsilon_N + \|\mathbf{w}\|^2\right)}, \tag{A1}$$

where $\mathcal{A}_N = \frac{\Gamma\left(\frac{N+1}{2}\right)}{\varepsilon_N^{\frac{N-1}{2}} 2^{\frac{N}{2}} \pi^{N+\frac{1}{2}} \sqrt{|\mathbf{R}|}}$ is a normalization constant. This integral is also called the partition function of the system in statistical physics since it sums up the contributions of all possible states to the evidence. To untangle the angular and radial degrees of freedom, we

apply the following identity for any $\alpha > 0$ and $\beta > 0$ to the prior:

$$\alpha^{-\beta} = \frac{1}{\Gamma(\beta)} \int_{-\infty}^{\infty} \mathrm{d}t \, e^{-\alpha e^t + \beta t} \,. \tag{A2}$$

Additionally assuming here that the observation operator is linear, we obtain:

$$p(\mathbf{y}) = \mathcal{B}_N \int_{\mathbb{R}^{N+1}} \mathrm{d}\mathbf{w}\mathrm{d}t \, e^{-\frac{1}{2}\|\boldsymbol{\delta} - \mathbf{Y}\mathbf{w}\|^2_{\mathbf{R}} - \frac{1}{2}e^t\|\mathbf{w}\|^2 - \frac{1}{2}e^t\varepsilon_N + \frac{N+1}{2}t} \,, \tag{A3}$$

where $\mathcal{B}_N = \frac{2^{\frac{N+1}{2}}}{\Gamma\left(\frac{N+1}{2}\right)}\mathcal{A}_N$. The main contribution to the evidence can be estimated by using the Laplace method to estimate this integral. Let us denote $\mathcal{L}(\mathbf{w}, t)$ minus the argument of the exponential in the integrant. If the saddle-point of $\mathcal{L}(\mathbf{w}, t)$ is $(\mathbf{w}_\star, t_\star)$, and if its Hessian at the saddle-point is $\mathcal{H}_{\mathbf{w},t}(\mathbf{w}_\star, t_\star)$, then an estimate of the evidence is (**?**):

$$p(\mathbf{y}) \simeq \mathcal{B}_N \frac{\sqrt{(2\pi)^{N+1}}}{|\mathcal{H}_{\mathbf{w},t}(\mathbf{w}_\star, t_\star)|} e^{-\mathcal{L}(\mathbf{w}_\star, t_\star)} \,. \tag{A4}$$

The normalization by the Hessian represents a correction due to Gaussian fluctuations of the variables $(\mathbf{w}, t)$ around the saddle-point. The saddle-point conditions are

$$\mathbf{w} = \left(\mathbf{Y}^\mathsf{T}\mathbf{R}^{-1}\mathbf{Y} + e^t\mathbf{I}_N\right)^{-1}\mathbf{Y}^\mathsf{T}\mathbf{R}^{-1}\boldsymbol{\delta} \,, \qquad e^t = \frac{N+1}{\varepsilon_N + \|\mathbf{w}\|^2} \,. \tag{A5}$$

which are equivalent to the dual EnKF-N saddle-point equations (**?**). The Hessian is

$$\mathcal{H}_{\mathbf{w},t}(\mathbf{w}_\star, t_\star) = \begin{bmatrix} \mathbf{Y}^\mathsf{T}\mathbf{R}^{-1}\mathbf{Y} + e^{t_\star}\mathbf{I}_N & e^{t_\star}\mathbf{w}^\star \\ e^{t_\star}\mathbf{w}^\star & \frac{N+1}{2} \end{bmatrix} . \tag{A6}$$

Hence, the integral is dominated by the saddle-point solution found in the dual EnKF-N derivation. It corresponds to a standard ETKF analysis with a prior correction by the $e^{t_\star}$ factor. Moreover, the fluctuations are due to the standard ETKF fluctuations $\mathbf{Y}^\mathsf{T}\mathbf{R}^{-1}\mathbf{Y} + e^{t_\star}\mathbf{I}_N$,

with additional corrections due to the radial degree of freedom. When computing a precision matrix $\mathcal{H}_{\mathbf{w}}$ for the variables $\mathbf{w}$ from the Hessian Eq. (A6) using the Schur complement, i.e. the precision on the $\mathbf{w}$ variables conditioned on the knowledge of $t_\star$, we find

$$\mathcal{H}_{\mathbf{w}}(\mathbf{w}_\star, t_\star) = \mathbf{Y}^\mathsf{T}\mathbf{R}^{-1}\mathbf{Y} + e^{t_\star}\mathbf{I}_N - \frac{2}{N+1}e^{2t^\star}\mathbf{w}_\star\mathbf{w}_\star^\mathsf{T}, \tag{A7}$$

which coincides with Eq. (24). This tells that the correction $-2(N+1)^{-1}\zeta^2\mathbf{w}_a\mathbf{w}_a^\mathsf{T}$ in Eq. (24) is due to the fluctuation of $\zeta(=e^t)$ and its coupling to the angular degrees of freedom.

Algorithm 1 Algorithm of the primal EnKF-N

Require: The forecast ensemble $\{\mathbf{x}_k\}_{k=1,\dots,N}$, the observations $\mathbf{y}$, the observation error covariance matrix $\mathbf{R}$, and $\mathbf{U}$ an orthogonal matrix satisfying $\mathbf{U1} = \mathbf{1}$.

1: Compute the mean $\overline{\mathbf{x}}$ and the perturbations $\mathbf{X}$ from $\{\mathbf{x}_k\}_{k=1,\dots,N}$, $\mathbf{Y} = \mathbf{HX}$

2: Find the argument of the minimum:
$\mathbf{w}_a = \underset{\mathbf{w}}{\operatorname{argmin}} \left[\|\mathbf{y} - H(\overline{\mathbf{x}} + \mathbf{Xw})\|^2_{\mathbf{R}} + (N+1)\ln\left(\varepsilon_N + \|\mathbf{w}\|^2\right)\right]$

3: Compute: $\mathcal{H}_a = \mathbf{Y}^\mathsf{T}\mathbf{R}^{-1}\mathbf{Y} + (N+1)\frac{\left(\varepsilon_N + \|\mathbf{w}_a\|^2\right)\mathbf{I}_N - 2\mathbf{w}_a\mathbf{w}_a^\mathsf{T}}{\left(\varepsilon_N + \|\mathbf{w}_a\|^2\right)^2}$

4: Compute $\mathbf{x}^a = \overline{\mathbf{x}} + \mathbf{Xw}_a$, $\mathbf{W}^a = \sqrt{N-1}\,[\mathcal{H}_a]^{-\frac{1}{2}}\,\mathbf{U}$

5: Compute $\mathbf{x}^a_k = \mathbf{x}^a + \mathbf{X}\,[\mathbf{W}^a]_k$

Algorithm 2 Algorithm of the dual EnKF-N

Require: The forecast ensemble $\{\mathbf{x}_k\}_{k=1,\dots,N}$, the observations $\mathbf{y}$, the observation error covariance matrix $\mathbf{R}$, and $\mathbf{U}$ an orthogonal matrix satisfying $\mathbf{U1} = \mathbf{1}$.

1: Compute the mean $\overline{\mathbf{x}}$ and the perturbations $\mathbf{X}$ from $\{\mathbf{x}_k\}_{k=1,\dots,N}$, $\mathbf{Y} = \mathbf{HX}$, $\boldsymbol{\delta} = \mathbf{y} - \mathbf{H}\overline{\mathbf{x}}$

2: Find the argument of the minimum:

$$\zeta_a = \underset{\zeta \in]0,(N+1)/\varepsilon_N]}{\operatorname{argmin}} \left[\boldsymbol{\delta}^{\mathsf{T}} \left(\mathbf{R} + \mathbf{Y}\zeta^{-1}\mathbf{Y}^{\mathsf{T}} \right)^{-1} \boldsymbol{\delta} + \varepsilon_N \zeta + (N+1) \ln \tfrac{N+1}{\zeta} - (N+1) \right]$$

3: Compute $\mathbf{w}_a = \left(\mathbf{Y}^{\mathsf{T}}\mathbf{R}^{-1}\mathbf{Y} + \zeta_a \mathbf{I}_N \right)^{-1} \mathbf{Y}^{\mathsf{T}}\mathbf{R}^{-1}\boldsymbol{\delta}$

4: Compute $\mathcal{H}_a = \mathbf{Y}^{\mathsf{T}}\mathbf{R}^{-1}\mathbf{Y} + \zeta_a \mathbf{I}_N - \frac{2\zeta_a^2}{N+1} \mathbf{w}_a \mathbf{w}_a^{\mathsf{T}}$

5: Compute $\mathbf{x}^a = \overline{\mathbf{x}} + \mathbf{X}\mathbf{w}_a$, $\mathbf{W}^a = \sqrt{N-1} \left[\mathcal{H}_a \right]^{-\frac{1}{2}} \mathbf{U}$

6: Compute $\mathbf{x}_k^a = \mathbf{x}^a + \mathbf{X}[\mathbf{W}^a]_k$

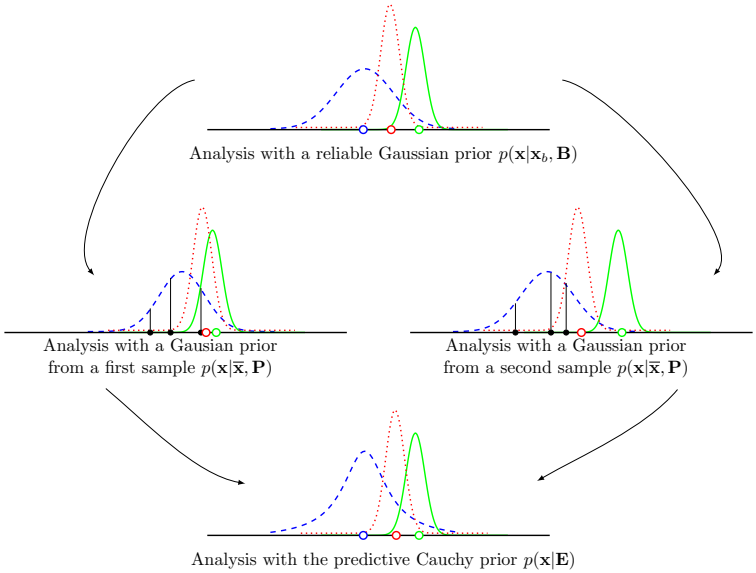


Figure 1. Schematic of the traditional standpoint on the analysis of the EnKF (top row), what it actually does using a Gaussian prior sampled from 3 particles (middle row), and using a predictive prior accounting for the uncertainty due to sampling (bottom row). The full green line represent the Gaussian observation error prior pdfs, the dashed blue lines represent the Gaussian/predictive priors if known, or estimated from an ensemble, or obtained from a marginalization over multiple potential errors statistics. The dotted red curves are the resulting analysis pdfs.

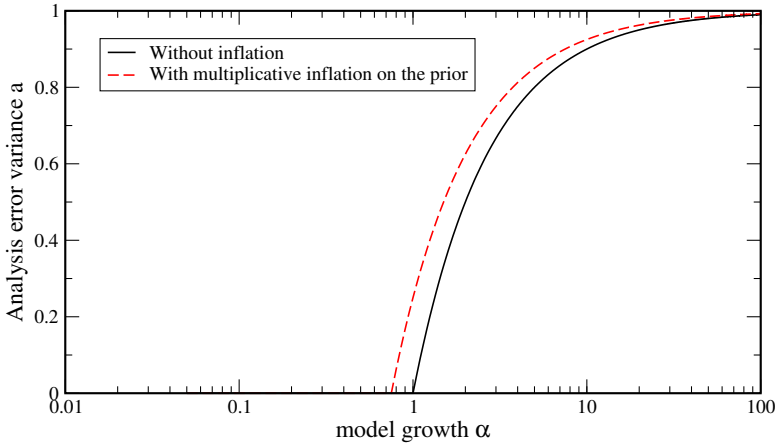


Figure 2. Analysis error variance when applying sequential data assimilation to $x_{k+1} = \alpha x_k$ with ($\zeta = 0.75$, dashed line) or without ($\zeta = 1$, full line) multiplicative inflation on the prior, as a function of the model growth α. We chose $r = 1$.

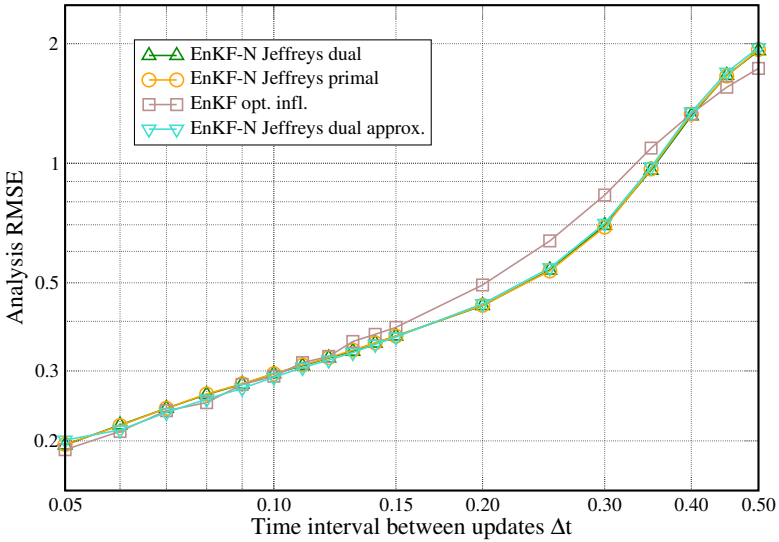


Figure 3. Average analysis RMSE for the primal EnKF-N, the dual EnKF-N, the approximate EnKF-N, and the EnKF with uniform optimally tuned inflation, applied to the Lorenz-95 model, as a function of the time step between updates. The finite-size EnKFs are based on Jeffreys' hyperprior.

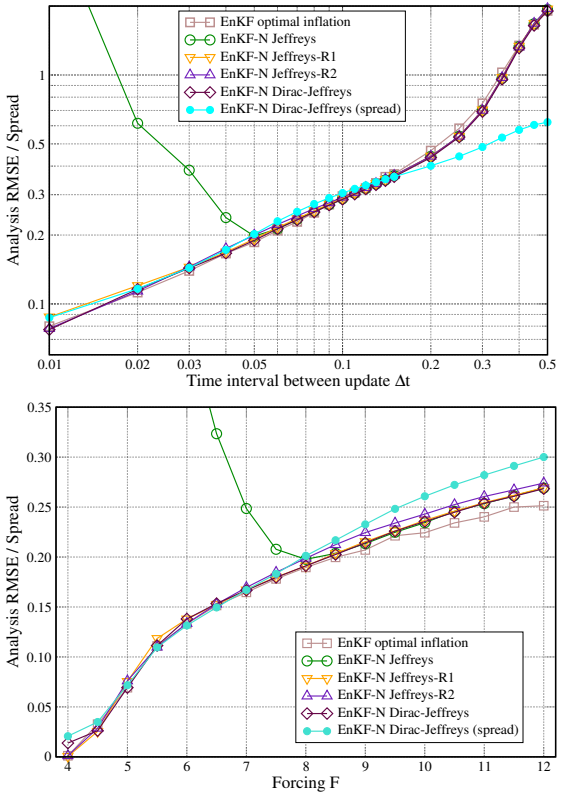


Figure 4. Average analysis RMSE for the EnKF-N with Jeffreys' hyperprior, with the EnKF-N based on the Dirac-Jeffreys' hyperprior, with the EnKF-N based on the Jeffreys' hyperprior but enforcing the schemes R1 or R2, and the EnKF with uniform optimally tuned inflation, applied to the Lorenz-95 model, as a function of the time step between update (top), and as a function of the forcing F of the Lorenz-95 model (bottom). The analysis ensemble spread of the EnKF-N based on the Dirac-Jeffreys' hyperprior is also shown.

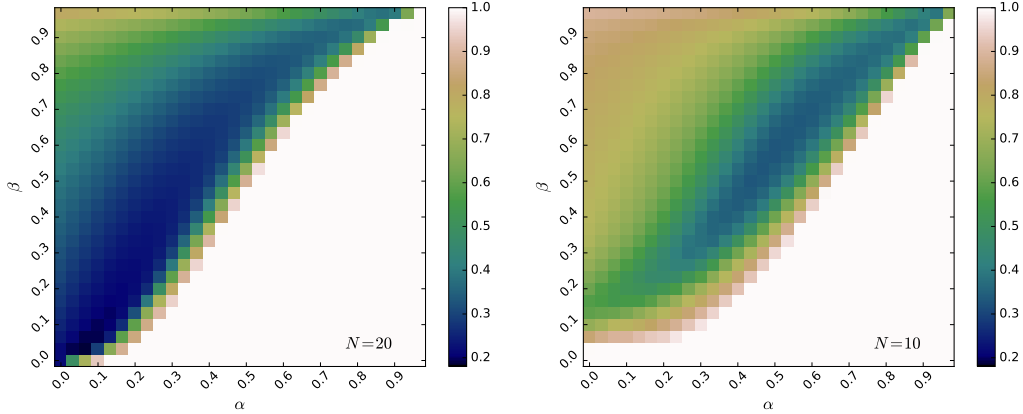


Figure 5. Average analysis RMSE as a function of (α, β) for the EnKF-N based on the IW hyperprior, without inflation nor enforced localization, for ensemble sizes of $N = 20$ (left) and of $N = 10$ (right). The RMSEs above 1, i.e. worse than an analysis based only on observations, are in white.