A Hierarchical Bayes Ensemble Kalman Filter

Michael Tsyrulnikov[a,*], Alexander Rakitko[a]

[a] *HydroMeteorological Research Center of Russia*

Abstract

A new ensemble filter that allows for the uncertainty in the prior distribution is proposed and tested. The filter relies on the conditional Gaussian distribution of the state given the filtering error covariances. The model-error and predictability-error covariance matrices are treated as random matrices and updated in a hierarchical Bayes scheme along with the state. The (hyper)prior distribution of the covariance matrices is assumed to be inverse Wishart. The new Hierarchical Bayes Ensemble Filter (HBEF) assimilates ensemble members as generalized observations and allows ordinary observations to influence the covariances. The actual probability distribution of the ensemble members is allowed to be different from the true one. An approximation that leads to a practicable analysis algorithm is proposed. The new filter is studied in numerical experiments with a doubly stochastic one-variable model of "truth". The model permits the assessment of the variance of the truth and of the true filtering error variance at each time instance. The HBEF is shown to outperform the EnKF and the HEnKF in a wide range of filtering regimes in terms of the accuracy of its estimates of the state and the accuracy of the estimated filtering error variance.

Keywords: Data assimilation, Random matrix, inverse Wishart distribution, Conditionally Gaussian model, Secondary filter, Observations

1. Introduction

Stochastic filtering and smoothing is a mathematical name for what is called in natural sciences data assimilation. Whenever we have three things: (1) an evolving system whose state is of interest to us, (2) an imperfect mathematical model of the system, and (3) incomplete and noise-contaminated observations, there is room for data assimilation. Currently, data assimilation techniques are extensively used in geophysics: meteorology, atmospheric chemistry, oceanography, land hydrology [e.g. 1], underground oil reservoir modeling [2], biogeochemistry [3], geomagnetism [4], and being explored in other areas like systems biology [5], epidemiology [6], ecology [7], and biophysics [8]. Data assimilation techniques have reached their most advanced level in meteorology.

To simplify the presentation of our technique, we confine ourselves to sequential discrete-time filtering, whose goal is to estimate the current state of the system given all present and past observations. This is a cycled procedure, each cycle consists of an observation update step (called in meteorology *analysis*) when current observations are assimilated, and a time update (forecast) step that propagates information on past observations forward in time.

1.1. Stochastic models of uncertainty

Virtually all advanced data assimilation methods rely on stochastic modeling of the underlying uncertainties in observations and in the forecast model. Historically, the first breakthrough in meteorological data assimilation was the introduction of the stochastic model of locally homogeneous and isotropic random fields and the least squares estimation approach based on correlation functions (optimal interpolation by Eliassen [9] and Gandin [10]). The second big advancement was the development of global multivariate forecast error *covariance models* no longer based on correlation functions but relying on more elaborate approaches like spectral and wavelet models, spatial filters, diffusion equations, etc. [11, 12, 13, 14, 15]; these (estimated "off-line") background-error models have been utilized in so-called *variational* data assimilation schemes [e.g. 11]. The third major invention so far was the Ensemble Kalman Filter (EnKF) by Evensen [16], in which the uncertainty of the system state is assumed to be Gaussian and represented by a Monte Carlo sample (ensemble), so that static forecast error covariance models are replaced by

*Corresponding author
Email address: mik.tsyrulnikov@gmail.com (Michael Tsyrulnikov)

dynamic and flow-dependent ensemble covariances. The EnKF has then developed into a wide variety of ensemble based techniques including ensemble-variational hybrids, e.g. [17, 18, 19].

There is another class of non-parametric Monte Carlo based filters called particle filters [e.g. 20]. They do not rely on the Gaussianity assumption and thus are better suited to tackle highly nonlinear problems, but the basic underlying idea of representing the unknown continuous probability density by a sum of a relatively small number of delta functions looks attractive for low-dimensional systems, whereas in high dimensions, its applicability remains to be convincingly shown. We do not consider particle filters in this paper.

In this research, we propose to retain a kind of Gaussianity because a *parametric* prior distribution has the advantage of bringing a lot of regularizing information in the vast areas of state space, where there are no nearby ensemble members. But we are going to relax the Gaussian assumption replacing it by a more general *conditionally* Gaussian model.

1.2. Uncertainty in the forecast error distribution

In the traditional EnKF, the forecast (background) uncertainty is characterized by the forecast error covariance matrix $\mathbf{B}$, which is estimated from the forecast ensemble. The problem is that this estimate cannot be precise, especially in high-dimensional applications of the EnKF, where the affordable ensemble size is much less than the dimensionality of state space. So, the forecast uncertainty in the EnKF is largely uncertain by itself [21, 22]. On the practical side, a common remedy here is a kind of regularization of the sample covariance matrix [e.g. 21]. But these techniques (of which the most widely used is covariance localization or tapering) are more or less *ad hoc* and have side effects, so a unifying paradigm to optimize the use of ensemble data in filtering is needed. On the theoretical side, there is an appropriate way to account for this uncertain uncertainty: hierarchical Bayes modeling (e.g. [23]).

1.3. Hierarchical Bayes estimation

In the classical non-Bayesian statistical paradigm, the state $\mathbf{x}$ (*parameter* in statistics) is considered to be non-random being subject of estimation from random forecast and random observations. Optimal interpolation is an example.

In the non-hierarchical Bayesian paradigm, both observations $\mathbf{y}$ and the state $\mathbf{x}$ are regarded as random. At the first level of the hierarchy, one specifies the observation likelihood $p(\mathbf{y}|\mathbf{x})$. As $\mathbf{x}$ is random, one introduces the second level of the hierarchy, the probability distribution of $\mathbf{x}$ that summarizes our knowledge of the state $\mathbf{x}$ before current observations $\mathbf{y}$ are taken into account, the *prior* distribution $p(\mathbf{x}|\boldsymbol{\vartheta})$. Here $\boldsymbol{\vartheta}$ is the non-random vector of parameters of the prior distribution (called hyperparameters). So, the non-hierarchical Bayesian modeling paradigm is, essentially, a two-level hierarchy ($\mathbf{y}|\mathbf{x}$ and $\mathbf{x}|\boldsymbol{\vartheta}$). In the analysis, the prior density $p(\mathbf{x}|\boldsymbol{\vartheta})$ is updated using the observation likelihood $p(\mathbf{y}|\mathbf{x})$ leading to the *posterior* density $p(\mathbf{x}|\mathbf{y})$. Note that the analysis step in the Kalman Filter can be viewed as an example of the two-level Bayesian hierarchy, in which the prior $\mathbf{x}|\mathbf{b},\mathbf{B}$ is the Gaussian distribution with the hyperparameter $\mathbf{b}$ being the predicted ensemble mean vector and the hyperparameter $\mathbf{B}$ the predicted ensemble covariance matrix. Variational assimilation can be regarded as a similar two-level Bayesian hierarchy with $\mathbf{b}$ being the deterministic forecast and $\mathbf{B}$ the pre-specified covariance matrix.

In the hierarchical Bayesian paradigm, not only observations and the state are random, the prior distribution is also assumed to be random (uncertain). Specifically, the hyperparameters $\boldsymbol{\vartheta}$ are assumed to be random variables having their own (hyper)prior distribution governed by hyperhyperparameters $\boldsymbol{\gamma}$. If $\boldsymbol{\gamma}$ are non-random, then we have a three-level hierarchy ($\mathbf{y}|\mathbf{x}$, $\mathbf{x}|\boldsymbol{\vartheta}$, and $\boldsymbol{\vartheta}|\boldsymbol{\gamma}$). The meaningful number of levels in the hierarchy depends on the observability of the higher-level hyperparameters: a hyperparameter is worth to be considered as random and subject of update if it is "reasonably" observed. We will rely in this study on a three-level hierarchy with the prior covariances as the random hyperparameter.

Historically, Le and Zidek [24] introduced uncertain covariance matrices in the static geostatistical non-ensemble estimation framework known as Kriging. Berliner [25] proposed to use the hierarchical Bayesian paradigm to account for uncertainties in parameters of error statistics used in data assimilation. Within the EnKF paradigm, Myrseth and Omre [26] added $\mathbf{b}$ and $\mathbf{B}$ to the traditional control vector assuming that $\mathbf{B}$ is the inverse Wishart distributed random matrix and the distributions $\mathbf{b}|\mathbf{B}$ and $\mathbf{x}|\mathbf{m},\mathbf{B}$ are multivariate Gaussian. Bocquet [27] took a different path and treated $\mathbf{b}$ and $\mathbf{B}$ as nuisance variables to be integrated out rather than updating them as the components of the control vector. His filter (developed further in [28, 29]) imposed prior distributions for random $\mathbf{b}$ and $\mathbf{B}$ in order to change the Gaussian prior of the state $\mathbf{x}$ to a more realistic continuous mixture of Gaussians.

In this study, we follow the general path of [26]. We propose to split $\mathbf{B}$ into the model error covariance matrix $\mathbf{Q}$ and the predictability error covariance matrix $\mathbf{P}$. The reason for such splitting is the fundamentally different nature of model errors (which are external to the filter) vs. predictability errors (which are

internal, i.e. determined by the filter). At the analysis step, following the hierarchical Bayes paradigm, we update $\mathbf{P}$ and $\mathbf{Q}$ along with the state $\mathbf{x}$ using both observation and ensemble data. Performance of the new filter is thoroughly tested in numerical experiments with a one-variable model. Note that the observation error covariance matrix is assumed to be precisely known in this study.

2. Background and notation

We start by outlining filtering techniques that have led to our approach, indicating those their aspects that are relevant for this paper. Thereby, we introduce the notation; the whole list of of main symbols can be found in AppendixD.

2.1. Bayesian filtering

The general Bayesian filtering paradigm assumes that unknown systems states $\mathbf{x}_k \in \mathbb{R}^n$ (where $k = 0, 1, \dots$ denotes the time instance and n the dimension of the state space) are random, subject to estimation from random observations $\mathbf{y}_{1:k} = (\mathbf{y}_1, \dots, \mathbf{y}_k)$. The true system states obey a Markov stochastic evolutionary model such that the *transition density* $p(\mathbf{x}_k|\mathbf{x}_{k-1})$ is available. Observations are related to the truth through the observation *likelihood* $p(\mathbf{y}_k|\mathbf{x}_k)$. The optimal filtering process consists in alternating forecast and analysis steps. At the forecast step the *predictive* density $p(\mathbf{x}_k|\mathbf{y}_{1:k-1})$ is computed. The goal of the analysis step is to compute the *filtering* density $p(\mathbf{x}_k|\mathbf{y}_{1:k})$.

At the analysis step, the predictive density is regarded as a prior density, which we denote by the superscript f (from "forecast"): $p^f(\mathbf{x}_k) = p(\mathbf{x}_k|\mathbf{y}_{1:k-1})$. The filtering density can similarly be viewed as the posterior density denoted by the superscript a (from "analysis"): $p^a(\mathbf{x}_k) = p(\mathbf{x}_k|\mathbf{y}_{1:k})$.

Direct computations of the predictive and filtering densities are feasible only for very low-dimensional problems. This difficulty can be alleviated if we turn to *linear* systems.

2.2. Linear observed system

The evolution of the truth is governed by the linear discrete in time stochastic dynamical system:

$$\mathbf{x}_k = \mathbf{F}_k \mathbf{x}_{k-1} + \boldsymbol{\varepsilon}_k, \tag{1}$$

where $\mathbf{F}_k$ the (linear) forecast operator, $\boldsymbol{\varepsilon}_k \sim \mathcal{N}(0, \mathbf{Q}_k)$ the model error, and $\mathbf{Q}_k$ the model error covariance matrix. Observations $\mathbf{y}_k$ are related to the state through the observation equation

$$\mathbf{y}_k = \mathbf{H}_k \mathbf{x}_k + \boldsymbol{\eta}_k, \tag{2}$$

where $\mathbf{H}_k$ is the (linear) observation operator, $\boldsymbol{\eta}_k \sim \mathcal{N}(0, \mathbf{R}_k)$ the observation error, and $\mathbf{R}_k$ the observation error covariance matrix.

2.3. Prior and posterior covariance matrices

Here we introduce the prior, posterior, and predictability covariance matrices, which will be extensively used throughout the paper. By $\mathbf{b}_k = \mathbb{E}\,\mathbf{x}_k|\mathbf{y}_{1:k-1}$, we denote the mean of the prior distribution and by

$$\mathbf{B}_k = \mathbb{E}\,[(\mathbf{x}_k - \mathbf{b}_k)(\mathbf{x}_k - \mathbf{b}_k)^\top|\mathbf{y}_{1:k-1}] \tag{3}$$

the prior covariance matrix. Similarly, $\mathbf{a}_k = \mathbb{E}\,\mathbf{x}_k|\mathbf{y}_{1:k}$ is the posterior mean and

$$\mathbf{A}_k = \mathbb{E}\,[(\mathbf{x}_k - \mathbf{a}_k)(\mathbf{x}_k - \mathbf{a}_k)^\top|\mathbf{y}_{1:k}] \tag{4}$$

the posterior covariance matrix. With the linear dynamics defined in Eq.(1), $\mathbf{b}_k$ and $\mathbf{B}_k$ can be related to $\mathbf{a}_{k-1}$ and $\mathbf{A}_{k-1}$:

$$\mathbf{b}_k = \mathbb{E}\,[\mathbf{F}_k \mathbf{x}_{k-1} + \boldsymbol{\varepsilon}_k|\mathbf{y}_{1:k-1}] = \mathbf{F}_k\,\mathbf{a}_{k-1} \tag{5}$$

and

$$\mathbf{B}_k = \mathbb{E}\,[(\mathbf{F}_k(\mathbf{x}_{k-1} - \mathbf{a}_{k-1}) + \boldsymbol{\varepsilon}_k) \cdot (\mathbf{F}_k(\mathbf{x}_{k-1} - \mathbf{a}_{k-1}) + \boldsymbol{\varepsilon}_k)^\top|\mathbf{y}_{1:k-1}] = \mathbf{P}_k + \mathbf{Q}_k, \tag{6}$$

where

$$\mathbf{P}_k = \mathbf{F}_k \mathbf{A}_{k-1} \mathbf{F}_k^\top \tag{7}$$

is the predictability error covariance matrix.

2.4. Kalman filter

For the linear system introduced in section 2.2, the mean-square optimal linear filter is the Kalman filter (KF). Its forecast step is

$$\mathbf{x}_k^f = \mathbf{F}_k \mathbf{x}_{k-1}^a, \tag{8}$$

where, we recall, the superscripts f and a stand for the forecast and analysis filter estimates, respectively. The analysis update is

$$\mathbf{x}_k^a = \mathbf{x}_k^f + \mathbf{K}_k(\mathbf{y}_k - \mathbf{H}_k \mathbf{x}_k^f), \tag{9}$$

where $\mathbf{K}_k$ is the so-called gain matrix:

$$\mathbf{K}_k = \mathbf{B}_k \mathbf{H}_k^\top (\mathbf{H}_k \mathbf{B}_k \mathbf{H}_k^\top + \mathbf{R}_k)^{-1}. \tag{10}$$

The posterior covariance matrix is

$$\mathbf{A}_k = (\mathbf{I} - \mathbf{K}_k \mathbf{H}_k)\mathbf{B}_k. \tag{11}$$

Note that Eqs.(8) and (9) constitute the so-called *primary filter* [30], in which the estimates of the *state* are updated. The primary filter uses the forecast error covariance matrix $\mathbf{B}_k$ computed in the *secondary filter*, which is comprised of Eqs.(10),(11), (6), and (7).

2.4.1. Remarks

1. The KF's forecast $\mathbf{x}_k^f$ and analysis $\mathbf{x}_k^a$ are exactly the prior and posterior means, therefore the above prior and posterior covariance matrices $\mathbf{B}_k$ and $\mathbf{A}_k$ are also the *error covariance* matrices of the filter's forecast and analysis, respectively.
2. The KF's secondary filter uses only observation operators and not observations themselves. As a consequence, the conditional covariance matrices $\mathbf{B}_k$, $\mathbf{A}_k$, and $\mathbf{P}_k$ coincide with their unconditional counterparts, $\underline{\mathbf{B}}_k$, $\underline{\mathbf{A}}_k$, and $\underline{\mathbf{P}}_k$ (this fact will be utilized below in section 4.3).
3. The KF produces forecast and analysis estimates $\mathbf{x}_k^f$ and $\mathbf{x}_k^a$ that are the best in the mean-square sense among all *linear* estimates. The KF estimates become optimal among *all* estimates if the involved error distributions are Gaussian. For highly non-Gaussian distributions, the KF can be significantly sub-optimal, so the (near) Gaussianity is implicitly assumed in the KF (this holds for the ensemble KF as well).

Unfortunately, the KF is still prohibitively expensive in high dimensions. This motivated the introduction and wide spread in geophysical and other applications of its Monte Carlo based approximation, the ensemble KF.

2.5. Ensemble Kalman filter (EnKF)

As compared with the KF, the EnKF replaces the most computer-time demanding step of forecasting $\mathbf{P}_k$ (via Eq.(7)) by its *estimation* from a (small) forecast ensemble. Members of this ensemble, $\mathbf{x}_k^{fe}(i)$ (where fe denotes the forecast ensemble, $i = 1, \ldots, N$, and N is the ensemble size) are generated by replacing the two uncertain quantities in Eq.(1), $\mathbf{x}_{k-1}$ and $\boldsymbol{\varepsilon}_k$, by their simulated counterparts, $\mathbf{x}_{k-1}^{ae}(i)$ and $\boldsymbol{\varepsilon}_k^e(i)$, respectively:

$$\mathbf{x}_k^{fe}(i) = \mathbf{F}_k \mathbf{x}_{k-1}^{ae}(i) + \boldsymbol{\varepsilon}_k^e(i). \tag{12}$$

Here the superscript ae stands for the analysis ensemble (see below in this subsection) and the superscript e for a simulated pseudo-random variable. Then, the sample $\{\mathbf{x}_k^{fe}(i)\}_{i=1}^N$ is used to compute the sample (ensemble) mean and the sample covariance matrix $\mathbf{S}_k$. The Kalman gain $\mathbf{K}_k$ is computed following Eq.(10), in which $\mathbf{B}_k$ is a somehow regularized $\mathbf{S}_k$ (normally, by applying variance inflation and spatial covariance localization, [e.g. 21]).

The analysis ensemble $\mathbf{X}_k^{ae} = \{\mathbf{x}_k^{ae}(i)\}$ is computed either deterministically by transforming the forecast ensemble [e.g. 31], or stochastically [e.g. 17]. In this study, we make use of the stochastic analysis ensemble generation technique, in which the observations are perturbed by adding their simulated observation errors $\boldsymbol{\eta}^e(i) \sim \mathcal{N}(\mathbf{0}, \mathbf{R})$ and then assimilated using $\mathbf{x}_k^{fe}(i)$ as the background:

$$\mathbf{x}_k^{ae}(i) = \mathbf{x}_k^{fe}(i) + \mathbf{K}_k(\mathbf{y}_k + \boldsymbol{\eta}^e(i) - \mathbf{H}\mathbf{x}^{fe}(i)). \tag{13}$$

Note that in practical applications, the forecast operator $\mathbf{F}_k$ is allowed to be nonlinear.

2.6. Methodological problems in the EnKF that can be alleviated using the hierarchical Bayes approach

1. In most EnKF applications, the prior covariance matrix is largely uncertain due to the insufficient ensemble size, which is not optimally accounted for. As a result, the filter's performance degrades.
2. In the EnKF analysis equations, there is no intrinsic feedback from observations to the forecast error covariances. The secondary filter is completely divorced from the primary one. This underuses the observational information (because observation-minus-forecast differences do contain information on forecast-error covariances) and requires external adaptation or manual tuning of the filter.

2.7. Hierarchical filters

By hierarchical filters, we mean those that aim at explicitly accounting for the uncertainties in the filter's error distributions using hierarchical Bayesian modeling.

2.7.1. Hierarchical Ensemble Kalman filter (HEnKF) by Myrseth and Omre [26]

Myrseth and Omre [26] were the first who used the Hierarchical Bayes approach to address the uncertainty in the forecast error covariance matrix within the EnKF. Here we outline their technique using our notation. To simplify the comparison of their filter with ours, we assume that the dynamics are linear and neglect the uncertainty in the prior mean vector $\mathbf{b}_k$ identifying it with the deterministic forecast $\mathbf{x}_k^f$. The HEnKF differs from the EnKF in the following respects.

(i) $\mathbf{B}_k$ is assumed to be a random matrix with the inverse Wishart prior distribution: $\mathbf{B}_k \sim \mathcal{IW}(\theta, \mathbf{B}_k^f)$, where θ is the scalar sharpness parameter and $\mathbf{B}_k^f$ the prior mean covariance matrix (see our AppendixA). $\mathbf{B}_k^f$ is postulated to be equal to the previous-cycle posterior mean covariance matrix.

(ii) The forecast ensemble members are assumed to be drawn from the Gaussian distribution $\mathcal{N}(\mathbf{b}_k, \mathbf{B}_k)$, where $\mathbf{B}_k$ is the true forecast error covariance matrix.

(iii) Having the inverse Wishart prior for $\mathbf{B}_k$ and independent Gaussian ensemble members drawn from $\mathcal{N}(\mathbf{b}_k, \mathbf{B}_k)$ implies that these ensemble members can be used to refine the prior distribution of $\mathbf{B}_k$. The respective posterior distribution of $\mathbf{B}_k$ is again inverse Wishart with the mean $\mathbf{B}_k^a$ equal to a linear combination of $\mathbf{B}_k^f$ and the ensemble covariance matrix $\mathbf{S}_k$ (see our AppendixB).

(iv) In generating the analysis ensemble members $\mathbf{x}_k^{ae}(i)$, the HEnKF perturbs not only observations (as in the EnKF) but also simultaneously the $\mathbf{B}_k$ matrix according to its posterior distribution.

The HEnKF was shown to outperform the EnKF in numerical experiments with simple low-order models for small ensemble sizes, as well as with an intermediate complexity model without model errors for a constant field [26].

2.7.2. EnKF-N "without intrinsic need for inflation" by Bocquet et al. [27, 28, 29]

In the EnKF-N, the prior mean and covariance matrices are assumed to be uncertain nuisance parameters with non-informative Jeffreys prior probability distributions. There is also a variant of the EnKF-N with an informative Normal-Inverse-Wishart prior for $(\mathbf{b}, \mathbf{B})$. With the Gaussian conditional distribution of the truth $\mathbf{x}|\mathbf{b}, \mathbf{B}$ and the perfect ensemble, the unconditional distribution of the truth given the forecast and the ensemble is analytically tractable and is proposed to replace, in the EnKF-N, the traditional Gaussian prior. The resulting analysis algorithm involves a non-quadratic minimization problem, which, as the authors argue, can be feasible in high-dimensional problems.

In numerical experiments with low-order models, the EnKF-N without a superimposed inflation was shown to be competitive with the EnKF with optimally tuned inflation. There were also indications that the EnKF-N can reduce the need in covariance localization.

2.7.3. Need for further research

Returning to the list of the EnKF's problems (section 2.6), we note that the HEnKF does address the first problem (the uncertainty in $\mathbf{B}_k$), but it does not address the second one (absence of feedback from observations to covariances in the EnKF). Next, assumption **(ii)** in section 2.7.1 is too optimistic, which will be discussed below in section 3.5 when we introduce our filter. Finally, the HEnKF is going to be very costly in high dimensions because of the need to sample from an inverse Wishart distribution. (Myrseth and Omre [26] note, though, that this computationally heavy sampling can be dropped, but, to the authors' knowledge, this opportunity has not yet been tested.)

The EnKF-N addresses both problems mentioned in section 2.6, but it relies on the assumption that forecast ensemble members are drawn from the same distribution as the truth (like the HEnKF

relies on its assumption **(ii)**). As we will argue in section 3.5, this cannot be guaranteed if background error covariances are uncertain. Besides, the EnKF-N has no memory in the covariances (as it does not explicitly update them). As we show below, updating and cycling the covariances can be useful.

Thus, both the HEnKF and the EnKF-N are important first contributions to the area of hierarchical filtering, but there is a lot of room in this area for further improvements and new approaches. This study presents one of them.

3. Hierarchical Bayes Ensemble (Kalman) Filter (HBEF)

3.1. Setup and idea

We formulate the HBEF for linear dynamics and linear observations, see Eqs.(1) and (2). Observation errors are Gaussian. Other settings come, mainly, from the formulation of conditions under which the EnKF actually works in geophysical applications:

1. The ensemble size is too small for sample covariance matrices to be accurate estimators.
2. The direct computation of the predictability error covariance matrix $\mathbf{P}_k$ as $\mathbf{F}_k\mathbf{A}_{k-1}\mathbf{F}_k^\top$ is unfeasible.
3. The model error covariance matrix $\mathbf{Q}_k$ is temporally variable and explicitly unknown.

We also hypothesize that

4. Conditionally on $\mathbf{Q}_k$, model errors are zero-mean Gaussian: $\boldsymbol{\varepsilon}_k|\mathbf{Q}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}_k)$.
5. We can draw independent pseudo-random samples from $\mathcal{N}(\mathbf{0}, \mathbf{Q}_k)$ with the true $\mathbf{Q}_k$.

Under these assumptions, the KF theory cannot be applied. In this research, we propose a theory and design a filter (the HBEF) that acknowledge in a more systematic way than this is done in the EnKF that the covariance matrices $\mathbf{Q}_k$ and $\mathbf{P}_k$ are substantially uncertain. We regard $\mathbf{Q}_k$ and $\mathbf{P}_k$ as additional (to the state $\mathbf{x}_k$) random matrix variate variables to be estimated along with the state. We represent both the prior and the posterior distributions hierarchically:

$$p(\mathbf{x}, \mathbf{P}, \mathbf{Q}) = p(\mathbf{P}, \mathbf{Q}) \cdot p(\mathbf{x}|\mathbf{P}, \mathbf{Q}) \tag{14}$$

and advance in time the two densities in the r.h.s. of this equation. Thereby the conditional density $p(\mathbf{x}|\mathbf{P}, \mathbf{Q})$ is shown below to remain Gaussian. This point is central to our approach. As for the marginal density $p(\mathbf{P}, \mathbf{Q})$, its exact evolution appears to be unavailable, so we introduce approximations to the prior, postulating it to be based on the inverse Wishart distribution at any assimilation cycle.

Actually, not only $\mathbf{Q}_k$ and $\mathbf{P}_k$ are uncertain, the prior conditional mean $\mathbf{b}_k$ is uncertain as well. But to simplify the presentation of our approach, we neglect the uncertainty in $\mathbf{b}_k$ and assume that $\mathbf{b}_k = \mathbf{x}_k^f$, where $\mathbf{x}_k^f$ is the deterministic forecast. For this reason, remark 1 in section 2.4.1 applies here.

A notational comment is in order. To avoid confusion of point *estimates* (produced by the filter) with their *true* counterparts, we mark the former with a superscript (f or a) or the tilde. E.g. $\mathbf{B}_k^a$ is the analysis point estimate of the true forecast error variance $\mathbf{B}_k$.

3.2. Observation and ensemble data to be assimilated

The HBEF aims to optimally assimilate not only conventional observations but also ensemble members. To estimate $\mathbf{Q}_k$ and $\mathbf{P}_k$, we split the forecast ensemble (computed on the interval between the time instances $k-1$ and k) $\mathbf{X}_k^{fe} = (\mathbf{x}_k^{fe}(1), \ldots, \mathbf{x}_k^{fe}(N))$ into two ensembles. The first one is the model error ensemble $\mathbf{X}_k^{me} = (\mathbf{x}_k^{me}(1), \ldots, \mathbf{x}_k^{me}(N))$, whose members are pseudo-random draws from the true distribution of model errors. The second ensemble is the predictability ensemble $\mathbf{X}_k^{pe} = (\mathbf{x}_k^{pe}(1), \ldots, \mathbf{x}_k^{pe}(N))$ defined below.

Note that this splitting of the forecast ensemble does *not* imply that the ensemble size is doubled. In the course of the traditional forecast ensemble, we suggest preventing model error perturbations from being added to the model fields while accumulating them in the model error ensemble members.

We denote the combined (observation and ensemble) data at the time k as $\mathbf{Y}_k = (\mathbf{y}_k, \mathbf{X}_k^{me}, \mathbf{X}_k^{pe})$. To assimilate these data, we need the respective likelihoods.

3.3. Observation likelihood

The Gaussianity of observation errors implies that the observation likelihood is, by definition,

$$p(\mathbf{y}_k|\mathbf{x}_k) \propto \mathrm{e}^{-\frac{1}{2}(\mathbf{y}_k - \mathbf{H}\mathbf{x}_k)^\top \mathbf{R}_k^{-1}(\mathbf{y}_k - \mathbf{H}\mathbf{x}_k)}. \tag{15}$$

3.4. Model error ensemble likelihood

From assumption 5 in section 3.1 and AppendixB, it follows that we can write down the likelihood of $\mathbf{Q}_k$ given the model error ensemble member $\mathbf{x}_k^{me}(i)$:

$$p(\mathbf{x}_k^{me}(i)|\mathbf{Q}_k) \propto \frac{1}{|\mathbf{Q}_k|^{1/2}} \mathrm{e}^{-\frac{1}{2}(\mathbf{x}_k^{me}(i))^\top \mathbf{Q}_k^{-1} \mathbf{x}_k^{me}(i)}, \tag{16}$$

where $|.|$ stands for the matrix determinant.

We emphasize that the existence of the likelihood $p(\mathbf{x}_k^{me}(i)|\mathbf{Q}_k)$, Eq.(16), implies that members of the model error ensemble $\mathbf{X}_k^{me}$ **can be viewed as observations on the true** $\mathbf{Q}_k$. This is because the likelihood provides the necessary relationship between the data we have ($\mathbf{x}_k^{me}(i)$ here) and the parameter we aim to estimate ($\mathbf{Q}_k$), see also AppendixB. For the whole ensemble, the likelihood becomes

$$p(\mathbf{X}_k^{me}|\mathbf{Q}_k) = \prod_{i=1}^{N} p(\mathbf{x}_k^{me}(i)|\mathbf{Q}_k) \propto |\mathbf{Q}|^{-\frac{N}{2}} \, \mathrm{e}^{-\frac{\mathrm{N}}{2} \mathrm{tr}(\mathbf{S}_k^{me} \mathbf{Q}_k^{-1})}, \tag{17}$$

where

$$\mathbf{S}_k^{me} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_k^{me}(i) \, \mathbf{x}_k^{me}(i)^\top \tag{18}$$

is the sample covariance matrix.

Remark. Equation (18) differs from the conventional sample covariance formula: the ensemble members are not centered by the ensemble mean and the sum is divided by N and not by $N-1$. These differences stem from our neglect of the uncertainty in $\mathbf{b}_k$. In practical problems, when we are not so sure about the mean, the conventional sample covariance matrix is to be preferred.

3.5. Predictability error ensemble likelihood

We note that both ordinary observations $\mathbf{y}_k$ and model error ensemble members $\mathbf{x}_k^{me}(i)$ are produced outside the filter. The likelihoods Eq.(15) and (16) relate $\mathbf{y}_k$ and $\mathbf{x}_k^{me}(i)$ to the variables ($\mathbf{x}_k$ and $\mathbf{Q}_k$, respectively), which are independent of the filter, too. So, the two likelihoods do influence the filter (they are, in fact, parts of its setup) but not vice versa.

This is in contrast to the predictability ensemble members $\mathbf{x}_k^{pe}(i)$, which are generated by the filter and are expected to be related to $\mathbf{P}_k$, which crucially depends on the filter's performance. For each k, the distribution of $\mathbf{x}_k^{pe}(i)$ and the true $\mathbf{P}_k$ are *determined* by the filter's performance. Therefore, we cannot *impose* any relationship between $\mathbf{x}_k^{pe}(i)$ and $\mathbf{P}_k$. We only can try to *reveal* this relationship.

In so doing, we notice that assumption 1 implies that the true covariances are unavailable to the filter, consequently, the filter cannot produce the predictability ensemble members $\mathbf{x}_k^{pe}(i)$ from a distribution with the true covariance matrix $\mathbf{P}_k$. This lack of relationship between $\mathbf{x}_k^{pe}(i)$ and $\mathbf{P}_k$ entails that, strictly speaking, there is no likelihood $p(\mathbf{x}_k^{pe}(i)|\mathbf{P}_k)$ and so the predictability ensemble members *cannot be regarded as observations on the true* $\mathbf{P}_k$. (For the same reason, members of the traditional forecast ensemble $\mathbf{X}_k^{fe}$ cannot be viewed as observations on $\mathbf{B}_k$ in any situation in which $\mathbf{B}_k$ is uncertain.) This important point is further illustrated below in sections 4.6 and 4.9 and discussed in section 5.2.

In order to come up with a mathematically sound way of extracting information on the true $\mathbf{P}_k$ contained in the predictability ensemble $\mathbf{X}_k^{pe}$, we use the following device. First, we postulate the existence of an auxiliary matrix variate random variable $\mathbf{\Pi}_k$ such that the predictability ensemble members $\mathbf{x}_k^{pe}(i)$ are Gaussian distributed with the known mean (identified with the deterministic forecast $\mathbf{x}_k^f$) and the covariance matrix $\mathbf{\Pi}_k$:

$$p(\mathbf{X}_k^{pe}|\mathbf{\Pi}_k) = \prod_{i=1}^{N} p(\mathbf{x}^{pe}(i)|\mathbf{\Pi}_k) \propto |\mathbf{\Pi}_k|^{-\frac{N}{2}} \, \mathrm{e}^{-\frac{\mathrm{N}}{2} \mathrm{tr}(\mathbf{S}_k^{pe} \mathbf{\Pi}_k^{-1})}, \tag{19}$$

where $\mathbf{S}_k^{pe}$ is the predictability ensemble sample covariance matrix:

$$\mathbf{S}_k^{pe} = \frac{1}{N} \sum_{i=1}^{N} (\mathbf{x}_k^{pe}(i) - \mathbf{x}_k^f) \, (\mathbf{x}_k^{me}(i) - \mathbf{x}_k^f)^\top. \tag{20}$$

Second, we assume that the true $\mathbf{P}_k$ has a (known) probability distribution related to $\mathbf{\Pi}_k$. Specifically, we assume that

$$\mathbf{P}_k|\mathbf{\Pi}_k \sim \mathcal{IW}(\theta, \mathbf{\Pi}_k), \tag{21}$$

where θ is the sharpness parameter (see AppendixA), which controls the spread of the distribution of $\mathbf{P}_k$ around its mean $\mathbf{\Pi}_k$ (the greater θ the smaller the spread).

Now, we observe that we have related $\mathbf{X}_k^{pe}$ to $\mathbf{\Pi}_k$ through the density $p(\mathbf{X}_k^{pe}|\mathbf{\Pi}_k)$, see Eq.(19), and $\mathbf{\Pi}_k$ to $\mathbf{P}_k$ through the density $p(\mathbf{P}_k|\mathbf{\Pi}_k)$, see Eq.(21). The resulting indirect relationship between $\mathbf{X}_k^{pe}$ and $\mathbf{P}_k$ will allow us to assimilate the former to update the latter.

Thus, we have the likelihoods for both ordinary observations and ensemble data. Next, we need the prior distributions.

3.6. Analysis: prior distribution

The analysis control vector comprises $\mathbf{x}$, $\mathbf{P}$, and $\mathbf{Q}$; we also have the auxiliary variable $\mathbf{\Pi}$ (a nuisance parameter). Note that here and elsewhere we drop the time index k whenever all variables in a given equation pertain to the same assimilation cycle k. We have to define a prior distribution (recall, denoted by the superscript f) for all these four variables combined. By the prior distribution, we mean the conditional distribution given all past assimilated data $\mathbf{Y}_{1:k-1}$. This conditioning is implicit throughout the paper in pdfs marked by the superscript f. We specify the joint prior hierarchically:

$$p^f(\mathbf{x},\mathbf{\Pi},\mathbf{P},\mathbf{Q}) = p^f(\mathbf{\Pi},\mathbf{P},\mathbf{Q})\, p^f(\mathbf{x}|\mathbf{\Pi},\mathbf{P},\mathbf{Q}) = p^f(\mathbf{Q})\, p^f(\mathbf{\Pi}|\mathbf{Q})\, p^f(\mathbf{P}|\mathbf{Q},\mathbf{\Pi})\, p(\mathbf{x}|\mathbf{P},\mathbf{Q}). \tag{22}$$

The key feature here (assumed at the start of filtering, i.e. at $k=1$, and proved below for $k>1$) is that the prior distribution of the state is *conditionally Gaussian* given $\mathbf{P},\mathbf{Q}$:

$$\mathbf{x}|\mathbf{P},\mathbf{Q} \sim \mathcal{N}(\mathbf{x}^f, \mathbf{B}=\mathbf{P}+\mathbf{Q}). \tag{23}$$

Now, consider the priors for the covariance matrices in Eq.(22). Starting with $p^f(\mathbf{Q})$, we hypothesize that there is a *sufficient statistic* $\mathbf{Q}^f$ and this sufficient statistic is produced by the secondary filter as an estimate of $\mathbf{Q}$ from past data, see section 3.9.3. Then, from sufficiency, the dependency on the past data in $p^f(\mathbf{Q}_k) \equiv p(\mathbf{Q}_k|\mathbf{Y}_{1:k-1})$ can be replaced by the dependency on $\mathbf{Q}^f$, so that $p^f(\mathbf{Q}) = p(\mathbf{Q}|\mathbf{Q}^f)$. Similarly, we postulate that $p^f(\mathbf{\Pi}|\mathbf{Q}) = p(\mathbf{\Pi}|\mathbf{P}^f)$, where $\mathbf{P}^f$ is also provided by the secondary filter, and that $p^f(\mathbf{P}|\mathbf{Q},\mathbf{\Pi}) = p(\mathbf{P}|\mathbf{\Pi})$, where the latter density is defined in Eq.(21). As a result, Eq.(22) writes

$$p^f(\mathbf{x},\mathbf{\Pi},\mathbf{P},\mathbf{Q}) = p(\mathbf{Q}|\mathbf{Q}^f)\, p(\mathbf{\Pi}|\mathbf{P}^f)\, p(\mathbf{P}|\mathbf{\Pi})\, p(\mathbf{x}|\mathbf{B}=\mathbf{P}+\mathbf{Q}). \tag{24}$$

Further, we model $p(\mathbf{Q}|\mathbf{Q}^f)$ and $p(\mathbf{\Pi}|\mathbf{P}^f)$ using the inverse Wishart distribution:

$$\mathbf{Q}|\mathbf{Q}^f \sim \mathcal{IW}(\chi,\mathbf{Q}^f) \qquad \text{and} \qquad \mathbf{\Pi}|\mathbf{P}^f \sim \mathcal{IW}(\phi,\mathbf{P}^f), \tag{25}$$

where χ and ϕ are the static sharpness parameters.

To summarize, the prior distribution is given in Eq.(24), where the first three densities in the r.h.s. are inverse Wishart and the last one is Gaussian. Prior to the analysis, we have the deterministic forecast $\mathbf{x}^f$ and the five parameters of the three (hyper)prior (inverse Wishart) distributions: $\mathbf{Q}^f$, $\mathbf{P}^f$, χ, ϕ, and θ. Now, we have to update the prior distribution using both ordinary and ensemble observations and come up with the posterior distribution.

3.6.1. Remarks

1. Conditional Gaussianity is a natural extension of the Gaussian assumption made in the KF and the EnKF and is crucial to the HBEF as it enables a computationally affordable analysis algorithm.
2. The choice of the inverse Wishart distribution is motivated by its *conjugacy* for the Gaussian likelihood [32, 33]. Conjugacy means that the posterior pdf belongs to the same distributional family as the prior. In our case, the inverse Wishart prior is not fully conjugate but it greatly simplifies derivations and makes the analysis equations partly analytically tractable.

3.7. Posterior

Multiplying the prior Eq.(24) by the three likelihoods, Eqs.(15), (17), and (19), we obtain the posterior for the extended control vector $(\mathbf{x},\mathbf{P},\mathbf{Q},\mathbf{\Pi})$:

$$\begin{aligned} p^a(\mathbf{x},\mathbf{P},\mathbf{Q},\mathbf{\Pi}) = p^f(\mathbf{x},\mathbf{P},\mathbf{Q},\mathbf{\Pi}\,|\,\mathbf{X}^{me},\mathbf{X}^{pe},\mathbf{y}) \propto \\ p^f(\mathbf{x},\mathbf{P},\mathbf{Q},\mathbf{\Pi}) \cdot p(\mathbf{y}|\mathbf{x}) \cdot p(\mathbf{X}^{me}|\mathbf{Q}) \cdot p(\mathbf{X}^{pe}|\mathbf{\Pi}) = \\ [p(\mathbf{Q}|\mathbf{Q}^f)\, p(\mathbf{X}^{me}|\mathbf{Q})] \cdot [p(\mathbf{\Pi}|\mathbf{P}^f)\, p(\mathbf{X}^{pe}|\mathbf{\Pi})] \cdot [p(\mathbf{P}|\mathbf{\Pi})] \cdot p(\mathbf{x}|\mathbf{B}=\mathbf{P}+\mathbf{Q}) \cdot p(\mathbf{y}|\mathbf{x}) \end{aligned} \tag{26}$$

Note that in densities marked by the superscript a, the dependency on the past and present data $\mathbf{Y}_{1:k}$ is implicit. Now, our goal is to transform Eq.(26) and reduce it to the required posterior $p^a(\mathbf{x}, \mathbf{P}, \mathbf{Q})$.

We start by simplifying the expressions in the first two brackets in the third line of Eq.(26). These are seen to be the two prior densities, $p(\mathbf{Q}|\mathbf{Q}^f)$ and $p(\mathbf{\Pi}|\mathbf{P}^f)$, updated by the respective ensemble data but not yet by ordinary observations. For this reason, we call them sub-posterior densities and denote by the tilde. For the inverse Wishart priors, Eq.(25), and the likelihoods Eqs.(17) and (19), the sub-posterior distributions are again inverse Wishart (see AppendixB):

$$\widetilde{p}(\mathbf{Q}) = p(\mathbf{Q}|\mathbf{Q}^f)\, p(\mathbf{X}^{me}|\mathbf{Q}) \sim \mathcal{IW}(\chi + N, \widetilde{\mathbf{Q}}), \tag{27}$$

$$\widetilde{p}(\mathbf{\Pi}) = p(\mathbf{\Pi}|\mathbf{P}^f)\, p(\mathbf{X}^{me}|\mathbf{\Pi}) \sim \mathcal{IW}(\phi + N, \widetilde{\mathbf{P}}), \tag{28}$$

with the mean values

$$\widetilde{\mathbf{Q}} = \frac{\chi \mathbf{Q}^f + N\mathbf{S}^{me}}{\chi + N} \qquad \text{and} \qquad \widetilde{\mathbf{P}} = \frac{\phi \mathbf{P}^f + N\mathbf{S}^{pe}}{\phi + N} \tag{29}$$

Next, we eliminate the nuisance matrix variate parameter $\mathbf{\Pi}$ from the posterior. The standard procedure in Bayesian statistics is to integrate $\mathbf{\Pi}$ out. But in our case we cannot do so analytically, instead we resort to the empirical Bayes approach [34] and replace in the posterior, Eq.(26), $\mathbf{\Pi}$ with its estimate $\widetilde{\mathbf{P}}$ (the mean of the sub-posterior distribution Eq.(28) defined in Eq.(29)). This allows us to get rid of the second bracket in Eq.(26) (because the expression there does not depend on the control vector $(\mathbf{x}, \mathbf{P}, \mathbf{Q})$ and no longer depends on $\mathbf{\Pi}$) and replace the third bracket by

$$\widetilde{p}(\mathbf{P}) = p(\mathbf{P}|\mathbf{\Pi} = \widetilde{\mathbf{P}}) \sim \mathcal{IW}(\theta, \widetilde{\mathbf{P}}) \tag{30}$$

(see Eq.(21)). As a result, we arrive at the following equation for the posterior density

$$p^a(\mathbf{x}, \mathbf{P}, \mathbf{Q}) \propto \widetilde{p}(\mathbf{P})\, \widetilde{p}(\mathbf{Q})\, [p(\mathbf{x}|\mathbf{B})\, p(\mathbf{y}|\mathbf{x})], \tag{31}$$

where $\mathbf{B} = \mathbf{P} + \mathbf{Q}$ and all the terms that contain the state $\mathbf{x}$ are placed inside the bracket.

To reduce the joint posterior Eq.(31) to the marginal posterior of $\mathbf{P}, \mathbf{Q}$ times the conditional posterior of $\mathbf{x}$ given $\mathbf{P}, \mathbf{Q}$ (i.e. to represent the posterior hierarchically), we should integrate $\mathbf{x}$ out of $p^a(\mathbf{x}, \mathbf{P}, \mathbf{Q})$. This can be easily done because both $\mathbf{x}$-dependent terms in the bracket are proportional to Gaussian pdfs w.r.t. $\mathbf{x}$, see Eqs.(23) and (15), and so is their product. To analytically integrate $p^a(\mathbf{x}, \mathbf{P}, \mathbf{Q})$ over $\mathbf{x}$, we complete the square in the exponent of the $p(\mathbf{x}|\mathbf{B})\, p(\mathbf{y}|\mathbf{x})$ expression (technical details are omitted) and take into account that the integral of a Gaussian pdf equals one, getting

$$l(\mathbf{B}|\mathbf{y}) = \int_{\mathbb{R}^n} p(\mathbf{x}|\mathbf{B})\, p(\mathbf{y}|\mathbf{x})\, \mathrm{d}\mathbf{x} \propto \frac{|\mathbf{A}|^{\frac{1}{2}}}{|\mathbf{B}|^{\frac{1}{2}}} \cdot \mathrm{e}^{-\frac{1}{2}(\mathbf{y}-\mathbf{H}\mathbf{x}^{\mathrm{f}})^\top (\mathbf{H}\mathbf{B}\mathbf{H}^\top + \mathbf{R})^{-1} (\mathbf{y}-\mathbf{H}\mathbf{x}^{\mathrm{f}})}, \tag{32}$$

where the matrix $\mathbf{A}$ is defined below in Eq.(37). It is worth noting that $l(\mathbf{B}|\mathbf{y})$ defined in Eq.(32) is, essentially, the *observation likelihood of the matrix* $\mathbf{B}$ defined as $p(\mathbf{y}|\mathbf{B})$: indeed, $p(\mathbf{y}|\mathbf{B}) = \int p(\mathbf{y}|\mathbf{x})\, p(\mathbf{x}|\mathbf{B})\, \mathrm{d}\mathbf{x}$, hence the notation $l(\mathbf{B}|\mathbf{y})$.

Now we obtain the final posterior

$$p^a(\mathbf{x}, \mathbf{P}, \mathbf{Q}) = p^a(\mathbf{P}, \mathbf{Q}) \cdot p^a(\mathbf{x}|\mathbf{P}, \mathbf{Q}). \tag{33}$$

Here, from Eqs.(31) and (32),

$$p^a(\mathbf{P}, \mathbf{Q}) = \int p^a(\mathbf{x}, \mathbf{P}, \mathbf{Q})\, \mathrm{d}\mathbf{x} \propto \widetilde{p}(\mathbf{P})\, \widetilde{p}(\mathbf{Q})\, l(\mathbf{P} + \mathbf{Q}|\mathbf{y}) \tag{34}$$

is the marginal posterior. Further, from Eqs.(31) and (34),

$$p^a(\mathbf{x}|\mathbf{P}, \mathbf{Q}) = \frac{p^a(\mathbf{x}, \mathbf{P}, \mathbf{Q})}{p^a(\mathbf{P}, \mathbf{Q})} \propto p(\mathbf{x}|\mathbf{B})\, p(\mathbf{y}|\mathbf{x}) \sim \mathcal{N}(\mathbf{m}^a(\mathbf{B}), \mathbf{A}(\mathbf{B})), \tag{35}$$

(where, we recall, $\mathbf{B} = \mathbf{P} + \mathbf{Q}$) is the conditional posterior. In Eq.(35), the proportionality $\propto$ is w.r.t. $\mathbf{x}$ (because $p^a(\mathbf{x}|\mathbf{P}, \mathbf{Q})$ is a probability density of $\mathbf{x}$),

$$\mathbf{m}^a(\mathbf{B}) = \mathbf{x}^f + \mathbf{A}\, \mathbf{H}^\top \mathbf{R}^{-1} (\mathbf{y} - \mathbf{H}\mathbf{x}^f) \tag{36}$$

is the conditional posterior expectation of $\mathbf{x}$, and

$$\mathbf{A} = \mathbf{A}(\mathbf{B}) = (\mathbf{B}^{-1} + \mathbf{H}^\top \mathbf{R}^{-1} \mathbf{H})^{-1} \tag{37}$$

is the conditional posterior (analysis error) covariance matrix.

3.7.1. Remarks

1. *Preservation of conditional Gaussianity in the analysis.* The posterior conditional distribution of the state $p^a(\mathbf{x}|\mathbf{P},\mathbf{Q})$, Eq.(35), appears to be Gaussian (coinciding with the traditional KF posterior given $\mathbf{B}=\mathbf{P}+\mathbf{Q}$, therefore Eqs.(36) and (37) are exactly the KF equations). This implies, importantly, that conditional Gaussianity of the state "survives" the analysis step.
2. Equation (37) implies that with the inverse Wishart priors for $\mathbf{P}$ and $\mathbf{Q}$, the posterior covariance matrix $\mathbf{A}$ is *not* inverse Wishart distributed. In principle, we could approximate the distribution of $\mathbf{A}$ by inverse Wishart, but we will not pursue this idea here.
3. The inverse Wishart priors for the covariance matrices significantly simplify the derivation of the posterior distribution, but at the expense of not solving the problem of noisy long-distance covariances. This implies that covariance localization should be applied to the ensemble covariances.
4. The linear combinations of the prior and ensemble covariance matrices in Eq.(29) resemble, on the one hand, the shrinkage estimator of a covariance matrix proposed by [35] and, on the other hand, the use of static and ensemble covariances in hybrid ensemble variational techniques [e.g. 18, 19].
5. Equation (32) shows that observations do influence the observation likelihood of $\mathbf{B}$ (through the innovation vector $\mathbf{y}-\mathbf{H}\mathbf{x}^f$), hence they do influence the marginal posterior of the covariances, $p^a(\mathbf{P},\mathbf{Q})$, see Eq. (34). This is the "mechanism" in the HBEF that provides the desired and absent in the KF, EnKF, and HEnKF feedback from observations to the forecast error covariances.
6. In the classical Bayesian filtering theory outlined in section 2.1, the predictive and filtering distributions are conditioned on *ordinary* observations $\mathbf{y}$. In the HBEF, we explicitly condition the posterior on both observation and ensemble data $\mathbf{Y}$. The two conditionings lead to different results, but this difference is an inevitable consequence of approximations due to the ensemble (Monte Carlo) approach. We will not distinguish between them in the sequel.

3.8. Analysis equations

Having the posterior $p^a(\mathbf{x},\mathbf{P},\mathbf{Q})$, see Eqs.(33)–(35), we now need equations to compute quantities needed for the next assimilation cycle. These are, first, point estimates of $\mathbf{x},\mathbf{P},\mathbf{Q}$ (which we call deterministic analyses) and second, the analysis ensemble $\mathbf{X}^{ae}$.

3.8.1. Posterior mean $\mathbf{x},\mathbf{P},\mathbf{Q}$

The deterministic analyses $\mathbf{x}^a,\mathbf{P}^a,\mathbf{Q}^a$ are defined as approximations to their respective posterior mean values. The latter are given, obviously, by the following equations

$$\mathbf{P}^a=\mathbb{E}\,\mathbf{P}=\int\int p^a(\mathbf{P},\mathbf{Q})\,\mathbf{P}\,\mathrm{d}\mathbf{P}\mathrm{d}\mathbf{Q},\qquad \mathbf{Q}^a=\mathbb{E}\,\mathbf{Q}=\int\int p^a(\mathbf{P},\mathbf{Q})\,\mathbf{Q}\,\mathrm{d}\mathbf{P}\mathrm{d}\mathbf{Q}, \tag{38}$$

$$\mathbf{B}^a=\mathbf{P}^a+\mathbf{Q}^a, \tag{39}$$

$$\mathbf{x}^a=\mathbb{E}\,\mathbf{x}=\mathbb{E}\,\mathbb{E}\,(\mathbf{x}|\mathbf{P},\mathbf{Q})=\mathbb{E}\,\mathbf{m}^a(\mathbf{P}+\mathbf{Q})=\int\int p^a(\mathbf{P},\mathbf{Q})\,\mathbf{m}^a(\mathbf{P}+\mathbf{Q})\,\mathrm{d}\mathbf{P}\mathrm{d}\mathbf{Q}, \tag{40}$$

where $\mathbf{m}^a(\mathbf{B})$ is given by Eq.(36), $p^a(\mathbf{P},\mathbf{Q})$ by Eq.(34), the expectation is over the posterior distribution, and the integration w.r.t. a *matrix* is explained in AppendixC.

The integrals in Eqs.(38) and (40) are not analytically tractable, so we introduce approximations. We present here two versions of the analysis equations: a Monte Carlo based and an empirical Bayes based (the simplest version).

3.8.2. Monte Carlo based deterministic analysis

Here, we approximate the integrals in Eqs.(38) and (40) using Monte Carlo simulation. More specifically, we employ the importance sampling technique [e.g. 36] with the proposal density $\widetilde{p}(\mathbf{P})\,\widetilde{p}(\mathbf{Q})$. Generating the Monte Carlo draws $\mathbf{P}^e(i)\sim\widetilde{p}(\mathbf{P})$, $\mathbf{Q}^e(i)\sim\widetilde{p}(\mathbf{Q})$ (where $i=1,\ldots,M$ and M is the size of the Monte Carlo sample), and computing $\mathbf{B}^e(i)=\mathbf{P}^e(i)+\mathbf{Q}^e(i)$, we obtain the estimates:

$$\mathbf{P}^a=\frac{\sum_{i=1}^M l[\mathbf{B}^e(i)|\mathbf{y}]\cdot\mathbf{P}^e(i)}{\sum_{i=1}^M l[\mathbf{B}^e(i)|\mathbf{y}]},\qquad \mathbf{Q}^a=\frac{\sum_{i=1}^M l[\mathbf{B}^e(i)|\mathbf{y}]\cdot\mathbf{Q}^e(i)}{\sum_{i=1}^M l[\mathbf{B}^e(i)|\mathbf{y}]}, \tag{41}$$

$$\mathbf{x}^a=\frac{\sum_{i=1}^M l[\mathbf{B}^e(i)|\mathbf{y}]\cdot\mathbf{m}^a[\mathbf{B}^e(i)]}{\sum_{i=1}^M l[\mathbf{B}^e(i)|\mathbf{y}]}. \tag{42}$$

Note that in view of Eq.(32), the resulting analysis is nonlinear in both $\mathbf{x}^f$ and $\mathbf{y}$.

Sampling from an inverse Wishart distribution can be expensive in high dimensions, so we propose, next, a cheap alternative.

3.8.3. The simplest deterministic analysis

Here, we neglect the $l(\mathbf{B}|\mathbf{y})$ term in Eq.(34) altogether, thus allowing, as in the HEnKF, no feedback from observations to the covariances. The reason for this neglect is that the information on $\mathbf{P}$ and $\mathbf{Q}$ that comes, first, from the background (prior) matrices $\mathbf{P}^f$ and $\mathbf{Q}^f$ and second, from the two ensembles $\mathbf{X}^{pe}$ and $\mathbf{X}^{me}$, summarized in the sub-posterior distributions $\widetilde{p}(\mathbf{P})$ and $\widetilde{p}(\mathbf{Q})$, is much richer than information on $\mathbf{P}$ and $\mathbf{Q}$ that comes from current observations through the $l(\mathbf{B}|\mathbf{y})$ term. Indeed, $\mathbf{P}^f$ and $\mathbf{Q}^f$ accumulate vast amounts of past (albeit aging) information on $\mathbf{P}$ and $\mathbf{Q}$. Model error ensemble members constitute, as we have discussed, N direct observations on $\mathbf{Q}$. Predictability ensemble members are N observations on $\mathbf{\Pi}$ (and so indirectly on $\mathbf{P}$). But there is only one set of current ordinary observations, that is, all current observations combined give rise to only one (very) noise contaminated observation on $\mathbf{HBH}^\top+\mathbf{R}$ (but note that with known $\mathbf{R}$, this is the only observation on the *true* $\mathbf{B}$). Therefore, we assume that in Eq.(34) $\widetilde{p}(\mathbf{P})$ and $\widetilde{p}(\mathbf{Q})$ are much more peaked w.r.t. $(\mathbf{P},\mathbf{Q})$ than $l(\mathbf{P}+\mathbf{Q}|\mathbf{y})$, so that the correction made to the sub-posterior by the relatively flat $l(\mathbf{P}+\mathbf{Q}|\mathbf{y})$ is rather small, and in the first approximation can be neglected. This simplification results in the marginal posterior

$$p^a(\mathbf{P},\mathbf{Q}) = \widetilde{p}(\mathbf{P})\cdot\widetilde{p}(\mathbf{Q}). \tag{43}$$

Both $\widetilde{p}(\mathbf{P})$ and $\widetilde{p}(\mathbf{Q})$ are inverse Wishart pdfs with the mean values $\widetilde{\mathbf{P}}$ and $\widetilde{\mathbf{Q}}$, respectively, so

$$\mathbf{P}^a = \widetilde{\mathbf{P}} \quad \text{and} \quad \mathbf{Q}^a = \widetilde{\mathbf{Q}}. \tag{44}$$

As for the deterministic analysis of the state, the integral in Eq.(40) remains analytically intractable, so we resort to the empirical Bayes estimate

$$\mathbf{x}^a = \mathbf{m}^a(\mathbf{B}^a), \tag{45}$$

which is just the KF's analysis with $\mathbf{B}^a = \mathbf{P}^a + \mathbf{Q}^a$ as the assumed forecast error covariance matrix.

3.8.4. Analysis ensemble

Here, the HBEF follows the stochastic EnKF, see Eq.(13), where $\mathbf{x}^{fe}(i) = \mathbf{x}^{pe}(i) + \mathbf{x}^{me}(i)$.

3.9. Forecast step

3.9.1. Primary filter

From Eq.(1) and assumption 4, we have

$$\mathbf{x}^f_k = \mathbb{E}\,\mathbf{x}_k|\mathbf{y}_{1:k-1} = \mathbf{F}_k\cdot\mathbb{E}\,\mathbf{x}_{k-1}|\mathbf{y}_{1:k-1} = \mathbf{F}_k\mathbf{x}^a_{k-1}, \tag{46}$$

which is essentially the KF's Eq.(8).

3.9.2. Preservation of conditional Gaussianity

Let us look at the basic state evolution Eq.(1). In that equation, $\boldsymbol{\varepsilon}_k|\mathbf{Q}_k$ is Gaussian and independent of $\mathbf{x}_{k-1}$. Further, as it follows from Eqs.(35) and (36), $\mathbf{x}_{k-1}$ is conditionally Gaussian given $\mathbf{A}_{k-1}$. Therefore, as $\mathbf{x}_k = \mathbf{F}_k\mathbf{x}_{k-1}+\boldsymbol{\varepsilon}_k$, we obtain that $\mathbf{x}_k|\mathbf{A}_{k-1},\mathbf{Q}_k$ is Gaussian. But if we examine the distribution in question $\mathbf{x}_k|\mathbf{P}_k,\mathbf{Q}_k$, we observe that with the additional technical assumption that $\mathbf{F}_k$ is invertible, conditioning $\mathbf{x}_k$ on $\mathbf{P}_k$ is equivalent to conditioning on $\mathbf{A}_{k-1}$ (in view of Eq.(7)), consequently, Gaussianity of $\mathbf{x}_k|\mathbf{A}_{k-1},\mathbf{Q}_k$ implies Gaussianity of $\mathbf{x}_k|\mathbf{P}_k,\mathbf{Q}_k$.

Thus, the basic HBEF's conditional Gaussianity assumption is preserved at the forecast step (as well as the analysis step, see remark 1 in section 3.7.1).

3.9.3. Secondary filter

At the forecast step, the secondary filter has to produce $\mathbf{P}^f$ and $\mathbf{Q}^f$ at the next assimilation cycle. We postulate persistence as the simplest evolution model for both $\mathbf{P}$ and $\mathbf{Q}$, so that

$$\mathbf{P}^f_k = \mathbf{P}^a_{k-1} \qquad \text{and} \qquad \mathbf{Q}^f_k = \mathbf{Q}^a_{k-1}. \tag{47}$$

3.9.4. Generation of the forecast ensembles

The predictability ensemble $\mathbf{X}^{pe}_k$ is generated by simply applying the forecast operator $\mathbf{F}_k$ to the analysis ensemble members $\mathbf{x}^{ae}_{k-1}(i)$, see section 3.8.4. The model error ensemble $\mathbf{X}^{me}_k$ is generated by directly sampling from the model error distribution $\mathcal{N}(\mathbf{0},\mathbf{Q}_k)$.

4. Numerical experiments with a one-variable model

In this proof-of-concept study, we tested the proposed filtering methodology in numerical experiments with a one-variable model model of "truth", so that we were able to draw justified conclusions on fundamental aspects of the HBEF. Note that in the case of the one-dimensional state space we follow the default multi-dimensional notation but without the bold face.

We compared the HBEF with

1. *The reference KF* that has access to the "true" model error variances Q_k and is allowed to directly compute $P_k = F_k A_{k-1} F_k^\top = F_k^2 A_{k-1}$.
2. *The stochastic EnKF* with the optimally tuned variance inflation factor.
3. *The Var*, the filter based on the analysis that uses the constant $\bar{B}$ (the abbreviation Var stands for the variational analysis, which normally uses the time-mean B).
4. *The HEnKF*, in which, we recall, the prior mean is excluded from the analysis control vector in order to make it comparable with the HBEF.

We evaluated the performance of each filter by two criteria. The main criterion reflects the accuracy of the *primary filter* measured by the root-mean-square error (RMSE) of the filter's deterministic analysis of the state. The deterministic analysis is defined to be the analysis computed with the deterministic forecast as the background and the forecast error covariance matrix provided by the respective filter. (With the forecast model described below and small ensemble sizes, the deterministic forecast appeared to work better than the ensemble mean.)

The second criterion represents the accuracy of the *secondary filter* in terms of the RMSE of the filter's estimates of B (for details, see section 4.8 below). Note that by the RMSE we understand the root-mean-square difference with the truth (the true B is defined below in section 4.3).

Besides the formal evaluation of the performance of the new filter, we also examined some other important aspects of the technique proposed. First, we verified that the conditional distribution of the state given the covariances is indeed Gaussian. Second, we confirmed that the forecast ensemble variances were often systematically different from the true error variances. Third, we evaluated the role of the feedback from observations to the covariances, which is present in the HBEF with the Monte Carlo based analysis and absent in the other filters.

To conduct the numerical experiments presented in this paper, we developed a software package in the R language. The code, which allows one to reproduce all the below experiments, and its description are available from `https://github.com/rakitko/hbef`.

4.1. Model of "truth"

We wish the time series of the truth to resemble the natural variability of geophysical, specifically, atmospheric fields like temperature or winds. We would also like to be able to change various aspects of the probability distribution of our modeling true time series, so that the model of truth be conveniently parametrized, with parameters controlling distinct features of the time series distribution.

4.1.1. Model equations

We start by postulating the basic discrete-time equation

$$x_k = F_k x_{k-1} + \sigma_k \varepsilon_k, \tag{48}$$

where x_k is the truth, F_k and σ_k are the scalars to be specified, and $\varepsilon_k \sim \mathcal{N}(0,1)$ is the driving discrete-time white noise. Given the sequences $\{F_k\}$ and $\{\sigma_k\}$, the solution to Eq.(48) is a Gaussian distributed non-stationary time series. The forecast operator F_k determines the time-dependent time scale of x_k or, in other words, controls the degree of stability of the system: forecast perturbations are amplified if $|F_k| > 1$ and damped otherwise. Both $\{F_k\}$ and $\{\sigma_k\}$ together determine the time-dependent variance V_k of the random process x_k. The noise multiplier σ_k is the model error standard deviation: $Q_k = \sigma_k^2$.

In nature, both the variance and the temporal length scale exhibit significant chaotic day-to-day changes. In order to simulate these changes (and thus to introduce intermittent non-stationarity in the process x_k), we let F_k and σ_k be random sequences by themselves, thus making our model *doubly stochastic* [37]. Specifically, let F_k be governed by the equation:

$$F_k - \bar{F} = \mu(F_{k-1} - \bar{F}) + \sigma_F \varepsilon_k^F, \tag{49}$$

where $\mu \in (0,1)$ is the scalar controlling the temporal length scale of the process F_k, σ_F is the scalar controlling, together with μ, the variance of F_k, ε_k^F is the driving $\mathcal{N}(0,1)$ white sequence, and $\bar{F}$ is the

mean level of the F_k process. Equation (49) is the classical first-order auto-regression and its solution F_k is a stationary random process.

Further, let σ_k (see Eq.(48)) be a log-Gaussian distributed (which prevents σ from attaining unrealistically close to zero values and makes it positive) stationary time series:

$$\sigma_k = \exp(\Sigma_k) \qquad \text{with} \qquad \Sigma_k = \varkappa \Sigma_{k-1} + \sigma_\Sigma \varepsilon_k^\Sigma. \tag{50}$$

Here, $\varkappa$, σ_Σ, and ε^Σ have the same meanings as their counterparts in Eq.(49): μ, σ_F, and ε^F, respectively. We finally assume that the three random sources in our model, namely, ε_k, ε_k^F, and ε_k^Σ are mutually independent. Note that the process x_k is conditionally, given $\{F_k\}$ and $\{\sigma_k\}$, Gaussian, whereas unconditionally, the distribution of x_k is non-Gaussian.

4.1.2. Comparison with the existing models of "truth"

The difference of our model from popular simple nonlinear deterministic models, e.g. the three-variable Lorenz model [38] or discrete-time maps used to test data assimilation techniques (say, logistic or Henon maps [39]), is that in the deterministic models instabilities are curbed by the nonlinearity, whereas in our model, these are limited by the time the random process $|F_k|$ remains above 1. The nonlinear deterministic models are chaotic whereas our model is stochastic.

One advantage of our model of truth is that it allows us to know not only the truth itself but also its time-specific variance V_k. Indeed, running the model Eq.(48) L times with independent realizations of the forcing process ε_k (and with the sequences F_k and σ_k fixed), we can easily assess V_k using square averaging of x_k over the L realizations.

Another advantage of the proposed model of truth is that it has as many as five independent parameters, $\bar{F}$, μ, $\varkappa$, σ_F, and σ_Σ, which can be independently changed and which control different important features of the stochastic dynamical system Eqs.(48)–(50). These features include magnitudes and time scales of the solution x_k, the model error variance Q_k, and the degree of stability of the system. Note that these aspects affect the behavior of not only the truth but also the filters we are going to test.

In addition, the linearity of our model of truth allows the use of the exact KF as an unbeatable benchmark, which again would not be possible with nonlinear deterministic models of truth.

Finally, we remark that the model defined by Eqs.(48)–(50) is, actually, *nonlinear* if regarded as a state-space model, i.e. if the model equations are written as a Markov model for the *vector* state variable $(x_k, F_k, \Sigma_k)^\top$. In particular, the solution x_k nonlinearly depends on the system noise ε_k^Σ.

4.1.3. Model parameters

To select the five internal parameters of the system in a physically meaningful way, we related them to the five external parameters: the mean time scale $\bar{\tau}_x$ of the process x_k, the time scales τ_F and τ_Σ of the processes F_k and Σ_k, the probability of the "local instability" $\pi = \mathsf{P}(|F_k| > 1)$, and the variability in the system-noise variance, which we quantify by $\mathsf{s.d.}\,\Sigma_k$, the standard deviation of Σ_k. We specified the external parameters and then calculated the internal ones; we omit the respective elementary formulas.

4.2. The "default" configuration of the experimental system

4.2.1. Model

In order to assign specific values to the five external parameters, we interpreted our system, Eqs.(48)–(50), as a very rough model of the Earth atmosphere. Specifically, we arbitrarily postulated that one time step in our system corresponds to 2 hours of time in the atmosphere. This implies that the weather-related characteristic time scale of 1 day in the atmosphere corresponds to the mean time scale $\bar{\tau}_x = 12$ time steps for our process x_k. This was the default value for $\bar{\tau}_x$ in the experiments described below. Further, for the "structural" time series F_k and Σ_k, we specified somewhat longer time scales, $\tau_F = \tau_\Sigma = 1.5\bar{\tau}_x$. Next, the default value of π was selected to be equal to 0.05 and $\mathsf{s.d.}\,\Sigma_k$ equal to 0.5—these two values gave rise to reasonable variability in the system. We also examined effects of deviations of π and $\mathsf{s.d.}\,\Sigma_k$ from their default values, as described below. The sensitivity of our results to the other parameters of the model appeared to be low.

4.2.2. Observations

We generated observations by applying Eq.(2) with $H_k = 1$ and $\eta_k \sim \mathcal{N}(0, R)$ (so that the observation error variance $R_k = R$ is constant in time). To select the default value of R, we specified the default ratio B/R. In meteorology, for most observations, this forecast error to observation error ratio is about 1, but only a fraction of all system's degrees of freedom is observed. In our scalar system, the only degree of freedom is observed, so, to mimic the sparsity of meteorological observations, we inflated the

observational noise and so reduced the default ratio B/R to be equal to 0.1. This appeared to roughly correspond to the default $\sqrt{R} = 9$. We also examined the effect of varying R: from the well observed case with $B/R \simeq 10$ to the poorly observed case with $B/R \simeq 0.01$.

4.2.3. Ensemble size

In real-world atmospheric applications N is usually several tens or hundreds whilst the dimensionality of the system n is up to billions. In our system $n = 1$, so we chose N to vary from 2 to 10 with the default value of $N = 5$.

4.2.4. Version and parameters of the HBEF

By default, the simplest version of the HBEF was used, see section 3.8.3. To complete the specification of the default HBEF, it remained to assign values to the three sharpness parameters χ, ϕ, and θ, which was done by manual tuning. The default respective values were $\chi = 5$, $\phi = 30$, and $\theta = 2$.

4.2.5. Other parameters of the experimental setup

In the EnKF, the tuned variance inflation factor was 1.005. In the HEnKF, the best sharpness parameter was found to be $\theta = 10$. If not stated otherwise, the below statistics were computed with the length of the time series (the number of assimilation cycles) equal to $2 \cdot 10^5$.

4.3. Estimation of the true prior variances B_k and signal variances V_k

For an in-depth exploration of the HBEF's secondary filter, knowledge of the *true* forecast error variance B_k is very welcome, just like exploring the behavior of a primary filter is facilitated if one has access to the truth x_k. In this section, we show that our experimental methodology enables the assessment of the true B_k as accurately as needed.

We start by noting that each filter produces estimates of its own forecast error (co)variances B_k. By construction, the (exact) KF produces forecast error variances that coincide with the true B_k. All approximate filters (including those considered in this study) can produce only estimates of the B_k, e.g. the HBEF produces the posterior estimate B_k^a, see Eq.(39). It is worth stressing that B_k produced by the KF cannot be used as a proxy to the true B_k of any other filter because the error (co)variances are filter specific. The true B_k for each filter and each k can be assessed as follows.

Recall that B_k is the conditional (given all assimilated data) forecast error variance. Two aspects are important for us here. (i) B_k is the forecast error *variance*; this suggests that it can be assessed by averaging squared errors of the deterministic forecast, $(x_k^f - x_k)^2$. (ii) B_k is the *conditional* error variance; this means that B_k depends on all assimilated so far observations, so in order to assess the true B_k, one has to perform the averaging of squared errors only for those trajectories of the truth and those observation errors that give rise to exactly (or even approximately) the same observations B_k is conditioned upon. This is a computationally unfeasible task even for a one-variable model. But the assessment of the *unconditional* forecast error covariance matrix $\underline{B}_k$ is feasible and parallels the estimation of the true variance V_k outlined in section 4.1.2.

Specifically, we performed L independent assimilation runs, in which the sequences of F_k and σ_k (as well as the sequence of the observation operators) were the same (thus preserving the specificity of each time instance), whereas the sequences of ε_k, η_k, and the random sources in the filters related to the generation of the analysis ensembles were simulated in each run randomly and independently from the other runs. Then we used the mean squared forecast error as a proxy to the true $\underline{B}_k$:

$$\hat{\underline{B}}_k = \langle (x_k^f - x_k)^2 \rangle, \tag{51}$$

where the angle brackets $\langle . \rangle$ denote averaging over the L runs. In our experiments $L = 500$.

As noted in remark 2 in section 2.4.1, the KF's conditional B_k does not depend on the assimilated observations at all and thus coincides with the unconditional $\underline{B}_k$. This is true for any non-adaptive EnKF, the HEnKF, and the simplest version of the HBEF as well. But for the HBEF with the Monte Carlo based analysis, where there is feedback from observations to the covariances, this is not exactly the case. However, as we discussed in section 3.8.3, the influence of observations on the posterior estimates of P_k and Q_k (and thus B_k) is relatively weak, so we used $\underline{B}_k$ as a proxy to B_k for the HBEF with the Monte Carlo based analysis as well. To simplify the notation, we do not distinguish (for any filter in question) between the true conditional variance B_k, the true unconditional variance $\underline{B}_k$, and the proxy $\hat{\underline{B}}_k$.

Thus, for any time k, we had at our disposal the variance of the truth V_k and each filter's true forecast error variance B_k.

4.3.1. Remarks

1. Our approach here is similar to that proposed in [40]. The difference is that in [40] the truth is deterministic (so that V_k cannot be assessed) and the forecast model is stochastic, whereas our model assumes that the truth is stochastic whilst the forecast model is deterministic.
2. In order to avoid confusion with the filters' internal *estimates* of B_k (e.g. B_k^a), we use the terms *assessment* or *proxy* to refer to $\hat{\underline{B}}_k$, which externally evaluates the actual performance of the filter using the access to the truth.
3. All numerical experiments presented in this paper were carried out with one and the same arbitrarily selected realization of the structural time series F_k and σ_k, so that for any k, the signal variance V_k is the same for all plots below. This holds also for any filter's true forecast error variance B_k, facilitating comparison of the different plots.

4.4. Model's behavior

Figure 1 displays typical time series segments of F_k and σ_k, as well as of the two true variances, V_k and B_k, for the default configuration of the HBEF. One can see that the variance V_k of the signal x_k can vary in time by as much as some two orders of magnitude, so the process x_k was significantly non-stationary, as it is the case, say, in meteorology. One can also observe that the system-noise standard deviation σ_k was correlated with both V_k and B_k (which is not surprising). Correlation between F_k and both V_k and B_k was also positive but lower. Both V_k and B_k tended to be high when both $|F_k|$ and σ_k are high (low predictability events), and low when both $|F_k|$ and σ_k are low (high predictability regimes). The light (pink) vertical strips indicate events associated with $F_k > 1$: during these events, model errors were amplified giving rise to high V_k and B_k. The dark (blue) vertical strips indicate events when $|F_k|$ was relatively low: during these events, the errors are seen to be damped. In general, the model behaved as expected.

4.5. Verifying the conditional Gaussianity of the state given (x^f, B)

From the equation

$$x|x^f, P, Q \sim \mathcal{N}(x^f, B = P + Q), \tag{52}$$

it is obvious that $x_k|x_k^f, B_k$ is Gaussian if and only if so is $x_k - x_k^f|B_k$. With the true x_k and B_k in hand, we were able to verify if indeed $x_k - x_k^f \,|\, B_k \sim \mathcal{N}(0, B_k)$. Fig.2(left) presents the respective q-q (quantile-quantile) plots. (Note that for a Gaussian density, the q-q plot is a straight line, with the slope proportional to the standard deviation of the empirical distribution.) One can see that $p(x_k - x_k^f|B_k)$ can indeed be very well approximated by a Gaussian density for low, medium, and high values of B_k (the three curves in Fig.2(left)). In contrast, the *unconditional* density $p(x_k - x_k^f)$ is significantly non-Gaussian, see Fig.2(right). So, the conditional Gaussianity of the state's prior distribution is confirmed in our numerical experiments.

4.6. The forecast ensemble members are **not** *drawn from the same distribution as the truth*

Here, we explore the actual probability distribution of the forecast ensemble members at any given time k. We demonstrate that for both the EnKF and the HBEF, the variance of this distribution is often substantially biased with respect to the respective true error variance.

We start by stating that in a single data assimilation run, we cannot find out from which probability distribution the forecast ensemble members at time k are drawn (because the ensemble size is small, see assumption 1 in section 3.1). But, following section 4.3, for each filter, we had at our disposal a number of assimilation runs that share the sequence of B_k. Then, **if** in each assimilation run, the forecast ensemble members were drawn from the distribution with the variance B_k (the "null hypothesis"), we would have $\mathbb{E}\, S_k = B_k$, where the expectation is over the population of independent assimilation runs. To check if this latter equality actually holds, we estimated $\mathbb{E}\, S_k$ as the sample mean $\langle S_k \rangle$ for each k separately using the sample of L assimilation runs.

The resulting time series of the *biases* $\langle S_k \rangle - B_k$ for the EnKF and the HBEF are displayed in Fig.3 (the two lower curves) along with their respective 95% bootstrap confidence intervals. The true error variances themselves B_k are also shown in Fig.3 (the two upper curves) to give an impression of the relative magnitude of the biases in $\langle S_k \rangle$.

One can see that that the biases in the ensemble variances were significantly non-zero when the true error variance was relatively large. For the EnKF, the deviation of $\mathbb{E}\, S_k$ from B_k sometimes reached 50% of B_k. For the HBEF, the biases were less but still significant. In the small forecast error regimes, the biases became insignificant. It is also interesting to notice that the large biases were mostly negative

implying that the filters were under-dispersive (despite the tuned variance inflation in the EnKF). Over a longer time window of 10^4 time steps, the confidence interval did *not* contain zero (i.e. the bias was significantly non-zero) 78% of time for the EnKF and 62% of time for the HBEF.

Thus, we have to reject the null hypothesis and admit that forecast ensemble members are often taken from a distribution which is significantly different from the true one. This warrants the introduction of the actual predictability ensemble variance Π_k that differs from the true variance P_k (see section 3.5).

The above results are worth comparing with those of Bishop and Satterfield [40], who found insignificant biases in the ensemble variances, see their Fig.2. One dissimilarity between their and our experiments was that an ensemble transform version of the EnKF was used in [40]. We employed the ensemble transform technique for both the EnKF and the HBEF and found that this led to some improvements but did not remove the biases in S_k (not shown). A plausible reason for the difference in the conclusions is that the system in [40] was much better observed than ours (they used R which was much less than the mean V_k, whereas in our study R was several times larger than the mean V_k).

4.7. Verifying the primary filters

Here, we examine the accuracy of the state estimates for the HBEF and the other filters (the Var, the EnKF, and the HEnKF). In the below figures, we display their analysis RMSEs with the reference-KF analysis RMSEs subtracted. Figure 4(top, left) shows the RMSEs as functions of the ensemble size N. One can see that the HBEF was by far the best filter. For small $N < 3$, the Var became more competitive then the EnKF and the HEnKF, but still worse than the HBEF.

Figure 4(top, right) shows the RMSEs as functions of $\sqrt{R}$. Again, the HBEF performed the best. Its relative superiority was especially substantial for the smaller values of $\sqrt{R}$. This can be explained by the prevalence of Q (which is more rigorously treated in the HBEF) over P (which is only sub-optimally treated in the HBEF) in this regime.

Figure 4(bottom, left) shows the RMSEs as functions of $\pi = \mathsf{P}(|F_k| > 1)$. One can see that the HBEF was uniformly and significantly better than the other filters. Note that all the filters gradually deteriorate w.r.t. the reference KF as the system becomes less stable (i.e. as π grows), which is meaningful because errors grow faster in a less stable system.

Figure 4(bottom, right) displays the RMSEs as functions of the degree of intermittency in the model error variance quantified by $\mathsf{s.d.}\,\Sigma$. We see that the HBEF was still uniformly and substantially better than the EnKF and the HEnKF. For the smallest values of $\mathsf{s.d.}\,\Sigma$, the Var became superior to the EnKF and the HEnKF and only slightly worse than the HBEF. The fact that the Var worked relatively better for the small $\mathsf{s.d.}\,\Sigma$ can be explained by noting that in this regime, when the variability in Q was low, the forecast error statistics were less variable and so the constant Var's $\bar{B}$ was relatively more suitable.

Thus, in terms of the analysis RMSEs, the HBEF demonstrated its overall superiority over the competing EnKF, HEnKF, and Var filters.

4.8. Verifying the secondary filters

Recall that the HBEF's secondary sub-filter produces the posterior estimate $B_k^a = P_k^a + Q_k^a$ of its true forecast error variance B_k. In this section, we examine the errors $B_k^a - B_k$, where B_k is assessed following section 4.3.

We compare the performance of the HBEF's secondary filter with the other filters, excluding the exact KF, whose secondary filter has no error by construction. Each of the approximate filters yields or postulates an estimate of its own B_k. Specifically, the HEnKF produces B_k^a as described in item (iii) in section 2.7.1. In the Var, the constant $\bar{B}$ is used as an estimate of B_k, so we associate $\bar{B}$ with B_k^a. Similarly, we identify the EnKF's inflated ensemble variance S_k with its B_k^a.

Having the true B_k for each filter, we computed the RMSE in B_k^a using averaging over the L independent assimilation runs as $\Delta_k = \sqrt{\langle (B_k^a - B_k)^2 \rangle}$. The resulting Δ_k for the HBEF and the EnKF are depicted in Fig.5, where the almost uniform and substantial superiority of the HBEF is evident.

Having square averaged Δ_k over time, we obtained the time mean RMSEs in B_k^a. In a similar way we computed the biases in B_k^a. The results of an experiment with 10^4 time steps are collected in Table 1, where it is seen that the HBEF was much more accurate in estimating its B_k than the Var, the EnKF, and the HEnKF in estimating their respective true forecast error variances.

4.9. Role of feedback from observations to forecast error covariances

The HBEF with the Monte Carlo based analysis (section 3.8.2) provides an optimized way to utilize observations in updating P and Q. In the default setup, this capability did not lead to any improvement in the performance scores (not shown), but it became significant when the filter's model error variance was *misspecified*.

Table 1: Accuracy of the filters' estimates of their own forecast error variance B_k

Filter	Error bias $\mathsf{mean}\,(B_k^a - B_k)$	RMSE $\mathsf{rms}\,(B_k^a - B_k)$	Mean true B $\mathsf{mean}\,(B_k)$
Var	-0.9	6.5	7.6
EnKF	-1.4	6.2	7.5
HEnKF	-1.8	4.4	7.2
HBEF	-0.5	3.2	7.0

Specifically, we let all the filters (including the KF) "assume" that the model error variance Q_k equals the true one multiplied by the distortion coefficient $q_{distort}$. For several values of $q_{distort}$ in the range from 1/16 to 16, we computed the RMSEs of the analyses of the state for all filters and plotted the results in Fig.6. In the HBEF with the Monte Carlo based analysis, the size of the Monte Carlo sample was $M = 100$, see Eqs.(41)–(42). To make the effect more pronounced, the observation error standard deviation was reduced to $\sqrt{R} = 1$.

From Fig.6 one can see that the overall performance of the HBEF with the Monte Carlo based analysis was better than that of the other filters, including, we emphasize, the (now, inexact) KF. The observations-to-covariances feedback present in the Monte Carlo based HBEF (and absent in the other filters) appeared especially useful for $q_{distort} < 1$. The improvement was bigger for $q_{distort} < 1$ than for $q_{distort} > 1$ because an underestimation of the forecast error covariances is potentially more problematic for any filter. Indeed, the overconfidence in the forecast leads to an underuse of observations and in extreme cases can even lead to filter divergence. This is why the settings with $q_{distort} < 1$ leaved more room for improvement, particularly due to the feedback from observations to the covariances.

Another interesting conclusion can be drawn from comparing the Monte Carlo based version of the HBEF with the optimally tuned parameter θ (asterisks in Fig.6) and the same version of the HBEF but with $\theta = \infty$ (crosses). Recall that θ controls the difference between the variance Π of the distribution of the predictability ensemble members and the true variance P. In the setting with $\theta = \infty$, the HBEF "assumes" that $\Pi = P$. Figure 6 clearly shows that it was indeed beneficial to get away from the traditional assumption $\Pi = P$. This again justifies our suggestion (see section 3.5) to allow the ensemble distribution to be different from the true one.

5. Discussion

5.1. Comparison with other approaches

The HBEF has two immediate predecessors, the HEnKF [26] and the EnKF-N [27, 29]. The HBEF differs from the HEnKF in the following aspects. First, in the HBEF we treat $\mathbf{Q}$ and $\mathbf{P}$ separately instead of using the total background error covariance matrix $\mathbf{B}$. Second, the HBEF's forecast step is based on the persistence forecasts for the posterior point estimates of $\mathbf{Q}$ and $\mathbf{P}$ instead of that for the analysis error covariance matrix. These two improvements have led to the substantially better performance of the HBEF as compared to the HEnKF. Another difference from the HEnKF is that the Monte Carlo based HBEF permits observations to influence $\mathbf{Q}$ and $\mathbf{P}$. Experimentally, this latter feature appeared to be beneficial only when $\mathbf{Q}$ was significantly misspecified, though.

As compared to the EnKF-N, which integrates $\mathbf{B}$ out of the prior distribution, the HBEF explicitly updates the covariance matrices. This introduces memory in the covariances, which, as we have seen in the numerical experiments, can be beneficial.

In contrast to both the HEnKF and the EnKF-N, the HBEF in its present formulation does not treat the uncertainty in the prior mean state vector (this may be worth exploring in the future). But the HBEF systematically treats the uncertainty in $\mathbf{Q}$, which was assumed to be known in [26] and equal to zero in [27, 28, 29].

5.2. Distribution of ensemble members

We have argued that if we try to account for (substantial) uncertainties in the covariances, then we cannot at the same time keep assuming that predictability ensemble members or traditional forecast ensemble members are drawn from the *true* prior conditional distribution. This theoretical conclusion is confirmed in the numerical experiments (section 4.6), which show that the filter often produces significantly *biased* estimates of the (time specific) true covariances. We have proposed to account for this

additional uncertainty by hypothesizing that the predictability ensemble members are actually drawn from a Gaussian distribution with an unknown covariance matrix $\mathbf{\Pi}_k$, which is different from the true $\mathbf{P}_k$ but related to it through a known probability distribution $p(\mathbf{P}_k|\mathbf{\Pi}_k)$. This device proved to be beneficial in the experiments with the misspecified model error variance Q (section 4.9).

5.3. Restrictions of the proposed technique

First, the HBEF heavily relies on the *conditional* Gaussian prior distribution of the state. It is this assumption that greatly simplifies the analysis algorithm, but in a nonlinear context, it becomes an approximation, whose validity is to be verified.

Second, the HBEF makes use of the inverse Wishart prior distribution for the covariance matrices. There is no justification for this hypothesis other than partial analytical tractability of the resulting analysis equations, so other choices can be explored.

5.4. Practical applications

In order to apply the proposed technique to real-world high-dimensional problems, simplifications are needed because the $n \times n$ covariance matrices will be too large to be stored and handled. The computational burden can be reduced in different ways. Here is one of them. First, let the covariances to be defined on a coarse grid. Second, localize (taper) the covariances and store only non-zero covariance matrix entries. Third, use the simplest version of the HBEF.

Another possibility is to fit a parametric covariance model to current covariances and impose persistence for the *parameters* of the model. In this case, the simplest version of the HBEF would become close to practical ensemble variational schemes, but with climatological covariances replaced by evolving recent-past-data based covariances.

In high dimensions, the persistence forecast for the covariances seems to be worth improving. Specifically, one may wish to somehow spatially smooth $\mathbf{P}^a_{k-1}$ and $\mathbf{Q}^a_{k-1}$ in Eq.(47)—because it is meaningful that smaller scales in $\mathbf{P}^a_{k-1}$ and $\mathbf{Q}^a_{k-1}$ have less chance to survive until the next assimilation cycle than larger scales. Another way to improve the empirical forecast of the covariance matrices is to introduce a kind of "regression to the mean" making use of the time mean covariances. This would imply that the HBEF would cover not only EnKF but also ensemble variational hybrids as a special case.

The ultimate goal with the HBEF will be to obtain effective covariance regularization as a by-product of the hierarchical analysis scheme without using any ad-hoc device (as it was proposed for the EnKF-N in [27] and partially tested in [29]).

6. Conclusions

The progress made in this study can be summarized as follows.

- We have acknowledged that in most applications, the EnKF works with: (i) the explicitly unknown and variable model error covariance matrix $\mathbf{Q}_k$, (ii) the partially known (through ensemble covariances) background error covariance matrix. Under these explicit restrictions, we have proposed a new Hierarchical Bayes Ensemble Filter (HBEF) that optimizes the use of observational and ensemble data by treating $\mathbf{Q}_k$ and the predictability error covariance matrix $\mathbf{P}_k$ as random matrices to be estimated in the analysis along with the state.

- We have shown that model error ensemble members can be treated as generalized observations on $\mathbf{Q}_k$. As for ensemble members that have contributions from predictability errors, we have argued that, strictly speaking, these cannot be regarded as observations on the true respective covariances. A technique to mitigate this problem has been proposed and tested, so that the predictability ensemble members are also assimilated in the HBEF as generalized observations.

- The prior and posterior distributions of the state are shown to remain conditionally (given $\mathbf{P}_k, \mathbf{Q}_k$) Gaussian provided that: (i) it is so at the start of the filtering, (ii) observation errors are Gaussian, (iii) the dynamics and the observation operators are linear, and (iv) model errors are conditionally Gaussian given $\mathbf{Q}_k$. Unconditionally, the prior and posterior distributions of the state are non-Gaussian.

- The HBEF is tested with a new one-variable doubly stochastic model of truth. The model has the advantage of providing the means to assess the instantaneous variance of the truth and the true filter's error variance.

- The availability of the true error variances has permitted us to experimentally prove that the forecast ensemble variances in both the EnKF and the HBEF are often significantly biased with respect to the true variances.
- A (Monte Carlo based) version of the HBEF is shown to benefit from the feedback from observations to the covariances.
- The HBEF is found superior to the Var, the EnKF, and the HEnKF under most regimes of the system and most data assimilation setups.
- The simplest version of the HBEF is designed to be affordable for practical high-dimensional applications on existing computers.

7. Acknowledgments

The authors are very grateful to the two anonymous reviewers, whose valuable comments helped to significantly improve the manuscript.

References

[1] W. Lahoz, B. Khattatov, R. Menard, Data assimilation, Springer, 2010.

[2] D. Oliver, Y. Zhang, H. Phale, Y. Chen, Distributed parameter and state estimation in petroleum reservoirs, Comput. and Fluids 46 (12) (2011) 70–77.

[3] C. Trudinger, M. Raupach, P. Rayner, I. Enting, Using the Kalman filter for parameter estimation in biogeochemical models, Environmetrics 19 (8) (2008) 849–870.

[4] A. Fournier, G. Hulot, D. Jault, W. Kuang, A. Tangborn, N. Gillet, E. Canet, J. Aubert, F. Lhuillier, An introduction to data assimilation and predictability in geomagnetism, Space Sci. Rev. 155 (1-4) (2010) 247–291.

[5] R. Yoshida, M. Nagasaki, R. Yamaguchi, S. Imoto, S. Miyano, T. Higuchi, Bayesian learning of biological pathways on genomic data assimilation, Bioinformatics 24 (22) (2008) 2592–2601.

[6] C. Rhodes, T. D. Hollingsworth, Variational data assimilation with epidemic models, J. Theor. Biol. 258 (4) (2009) 591–602.

[7] S. Niu, Y. Luo, M. C. Dietze, T. F. Keenan, Z. Shi, J. Li, F. S. C. III, The role of data assimilation in predictive ecology, Ecosphere 5 (5) (2014) art65.

[8] D. Chapelle, M. Fragu, V. Mallet, P. Moireau, Fundamental principles of data assimilation underlying the Verdandi library: applications to biophysical model personalization within euHeart, Med. Biol. Eng. Comput. 51 (11) (2013) 1221–1233.

[9] A. Eliassen, Provisional report on calculation of spatial covariance and autocorrelation of the pressure field, Videnskaps-Akademiets Institutt for Vaer-og Klimaforskning, Oslo, Norway, Report No.5 (1954) 1–11.

[10] L. Gandin, Objective Analysis of Meteorological Fields, Gidrometizdat, Leningrad, 1963. Translated from Russian into English by the Israel Program for Scientific Translations, Jerusalem, 1965.

[11] F. Rabier, A. McNally, E. Andersson, P. Courtier, P. Unden, J. Eyre, A. Hollingsworth, F. Bouttier, The ECMWF implementation of three-dimensional variational assimilation (3D-Var). II: Structure functions, Q. J. Roy. Meteorol. Soc. 124 (550) (1998) 1809–1829.

[12] M. Fisher, Background error covariance modelling, Proc. ECMWF Semin. on recent developments in data assimilation for atmosphere and ocean, 8-12 September 2003 (2003) 45–64.

[13] A. Deckmyn, L. Berre, A wavelet approach to representing background error covariances in a limited area model, Mon. Weather Rev. 133 (5) (2005) 1279–1294.

[14] R. Purser, W. Wu, Numerical aspects of the application of recursive filters to variational statistical analysis. Part I: Spatially homogeneous and isotropic Gaussian covariances, Mon. Weather Rev. 131 (8) (2003) 1524–1535.

[15] A. Weaver, P. Courtier, Correlation modelling on the sphere using a generalized diffusion equation, Q. J. Roy. Meteorol. Soc. 127 (575) (2001) 1815–1846.

[16] G. Evensen, Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics, J. Geophys. Res. 99 (10) (1994) 10,143–10,162.

[17] P. L. Houtekamer, H. Mitchell, A sequential ensemble Kalman filter for atmospheric data assimilation, Mon. Weather Rev. 129 (1) (2001) 123–137.

[18] M. Buehner, J. Morneau, C. Charette, Four-dimensional ensemble-variational data assimilation for global deterministic weather prediction, Nonlin. Process. Geophys. 20 (5) (2013) 669–682.

[19] A. C. Lorenc, N. E. Bowler, A. M. Clayton, S. R. Pring, D. Fairbairn, Comparison of hybrid-4DEnVar and hybrid-4DVar data assimilation methods for global NWP, Mon. Weather Rev. 143 (2015) (2014) 212–229.

[20] P. van Leeuwen, Particle filtering in geophysical systems, Mon. Weather Rev. 137 (12) (2009) 4089–4114.

[21] R. Furrer, T. Bengtsson, Estimation of high-dimensional prior and posterior covariance matrices in Kalman filter variants, J. Multivar. Anal. 98 (2) (2007) 227–255.

[22] W. Sacher, P. Bartello, Sampling errors in ensemble Kalman filtering. part I: Theory, Mon. Weather Rev. 136 (8) (2008) 3035–3049.

[23] C. Robert, The Bayesian choice, Springer, 2007.

[24] N. D. Le, J. V. Zidek, Statistical analysis of environmental space-time processes, Springer, 2006.

[25] L. M. Berliner, Hierarchical Bayesian time series models, in: Maximum entropy and Bayesian methods, Springer, 1996, pp. 15–22.

[26] I. Myrseth, H. Omre, Hierarchical ensemble Kalman filter, SPE Journal 15 (2) (2010) 569–580.

[27] M. Bocquet, Ensemble Kalman filtering without the intrinsic need for inflation, Nonlin. Process. Geophys. 18 (5) (2011) 735–750.

[28] M. Bocquet, P. Sakov, Combining inflation-free and iterative ensemble kalman filters for strongly nonlinear systems, Nonlin. Process. Geophys. 19 (3) (2012) 383–399.

[29] M. Bocquet, P. Raanes, A. Hannart, Expanding the validity of the ensemble kalman filter without the intrinsic need for inflation, Nonlin. Process. Geophys. 22 (6) (2015) 645–662.

[30] D. P. Dee, S. E. Cohn, A. Dalcher, M. Ghil, An efficient algorithm for estimating noise covariances in distributed systems, IEEE Trans. Autom. Control 30 (11) (1985) 1057–1065.

[31] M. K. Tippett, J. L. Anderson, C. H. Bishop, T. M. Hamill, J. S. Whitaker, Ensemble square root filters, Mon. Weather Rev. 131 (7) (2003) 1485–1490.

[32] T. Anderson, An introduction to multivariate statistical analysis, Wiley Interscience, 2003.

[33] A. Gelman, J. Carlin, H. Stern, D. Rubin, Bayesian data analysis, Chapman and Hall/CRC, 2004.

[34] B. P. Carlin, T. A. Louis, Bayes and empirical Bayes methods for data analysis, Chapman and Hall/CRC, 2000.

[35] O. Ledoit, M. Wolf, A well-conditioned estimator for large-dimensional covariance matrices, J. Multivar. Anal. 88 (2) (2004) 365–411.

[36] D. Kroese, T. Taimre, Z. Botev, Handbook of Monte Carlo methods, Wiley, 2011.

[37] D. Tjøstheim, Some doubly stochastic time series models, J. Time Ser. Anal. 7 (1) (1986) 51–72.

[38] E. N. Lorenz, Deterministic nonperiodic flow, J. Atmos. Sci. 20 (2) (1963) 130–141.

[39] H. Du, L. A. Smith, Parameter estimation through ignorance, Physical Review E 86 (1) (2012) 016213.

[40] C. H. Bishop, E. A. Satterfield, Hidden error variance theory. Part I: Exposition and analytic model, Mon. Weather Rev. 141 (5) (2013) 1454–1468.

[41] A. K. Gupta, D. K. Nagar, Matrix variate distributions, CRC Press, 1999.

AppendixA. inverse Wishart distribution

In Bayesian statistics, the inverse Wishart distribution (e.g. [41, 32, 33]) is the standard choice for the prior distribution of a covariance matrix, because inverse Wishart is the so-called conjugate distribution for the Gaussian likelihood, e.g. [32, 33]. The inverse Wishart pdf is defined for symmetric matrices and is non-zero for positive definite ones:

$$p(\mathbf{Z}) \propto \frac{1}{|\mathbf{Z}|^{\frac{\nu+n+1}{2}}} \mathrm{e}^{-\frac{1}{2}\,\mathrm{tr}(\mathbf{Z}^{-1}\boldsymbol{\Sigma})}, \tag{A.1}$$

where $\nu > n+1$ is the so-called number of degrees of freedom (which controls the spread of the distribution: the greater ν, the less the spread) and $\boldsymbol{\Sigma}$ is the scaling positive definite matrix. Using the mean value $\bar{\mathbf{Z}} = \mathbb{E}\,\mathbf{Z} = \boldsymbol{\Sigma}/(\nu - n - 1)$ instead of the scaling matrix allows us to reparametrize Eq.(A.1) as

$$p(\mathbf{Z}) = p(\mathbf{Z}|\theta, \bar{\mathbf{Z}}) \propto \frac{1}{|\mathbf{Z}|^{\frac{\theta}{2}+n+1}} \mathrm{e}^{-\frac{\theta}{2}\,\mathrm{tr}(\mathbf{Z}^{-1}\bar{\mathbf{Z}})}, \tag{A.2}$$

where we have introduced a new scale parameter $\theta = \nu - n - 1 > 0$, which we call the *sharpness* parameter (the higher θ, the narrower the density). We symbolically write Eq.(A.2) as

$$\mathbf{Z} \sim \mathcal{IW}(\theta, \bar{\mathbf{Z}}). \tag{A.3}$$

We prefer our parametrization $(\theta, \bar{\mathbf{Z}})$ to the common one $(\nu, \boldsymbol{\Sigma})$ because $\bar{\mathbf{Z}}$ has the clear meaning of the (important) mean $\mathbf{Z}$ matrix. Summarizing, the inverse Wishart pdf has two parameters: the sharpness parameter θ (a scalar) and the mean $\bar{\mathbf{Z}}$ (a positive definite matrix).

AppendixB. Assimilation of conditionally Gaussian generalized observations in an update of their covariance matrix

Here, we outline, following e.g. [33], the procedure of assimilation of independent draws from the distribution $\mathcal{N}(\mathbf{m}, \mathbf{Z})$, where $\mathbf{m}$ is the known vector and $\mathbf{Z}$ the unknown random symmetric positive definite matrix, whose prior distribution is inverse Wishart with the density specified by Eq.(A.2).

Let us take a draw $\mathbf{x}^e(i)|\mathbf{Z} \sim \mathcal{N}(\mathbf{m}, \mathbf{Z})$, which we interpret as a member of an ensemble. Then, obviously,

$$p(\mathbf{x}^e(i)|\mathbf{Z}) \propto \frac{1}{|\mathbf{Z}|^{\frac{1}{2}}} \mathrm{e}^{-\frac{1}{2}(\mathbf{x}^{\mathrm{e}}(\mathrm{i})-\mathbf{m})^{\top}\mathbf{Z}^{-1}(\mathbf{x}^{\mathrm{e}}(\mathrm{i})-\mathbf{m})}. \tag{B.1}$$

We stress that Eq.(B.1) is nothing other than the likelihood of $\mathbf{Z}$ given the ensemble member $\mathbf{x}^e(i)$. Further, having the ensemble $\mathbf{X}^e = (\mathbf{x}^e(1), \ldots, \mathbf{x}^e(N))$ of N independent members all taken from $\mathcal{N}(\mathbf{m}, \mathbf{Z})$, we can write down the respective *ensemble likelihood* as the product of the partial likelihoods:

$$p(\mathbf{X}^e|\mathbf{Z}) \propto \frac{1}{|\mathbf{Z}|^{\frac{N}{2}}} \mathrm{e}^{-\frac{1}{2}\sum_{\mathrm{i}=1}^{\mathrm{N}}(\mathbf{x}^{\mathrm{e}}(\mathrm{i})-\mathbf{m})^{\top}\mathbf{Z}^{-1}(\mathbf{x}^{\mathrm{e}}(\mathrm{i})-\mathbf{m})} = \frac{1}{|\mathbf{Z}|^{\frac{N}{2}}} \mathrm{e}^{-\frac{N}{2}\,\mathrm{tr}(\mathbf{S}\mathbf{Z}^{-1})}, \tag{B.2}$$

where

$$\mathbf{S} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}^e(i) - \mathbf{m})\,(\mathbf{x}^e(i) - \mathbf{m})^{\top} \tag{B.3}$$

is the sample covariance matrix.

But having the likelihood $p(\mathbf{X}^e|\mathbf{Z})$ means that $\mathbf{X}^e$ (and its members $\mathbf{x}^e(i)$) can be regarded and treated as (generalized) *observations* on $\mathbf{Z}$. In particular, the ensemble can be *assimilated* in the standard way using the Bayes theorem. Indeed, having the prior pdf of $\mathbf{Z}$, Eq.(A.2), we obtain the posterior

$$p^a(\mathbf{Z}) \propto p(\mathbf{Z}|\theta, \bar{\mathbf{Z}}, \mathbf{X}^e) \propto p(\mathbf{Z}|\theta, \bar{\mathbf{Z}}) \cdot p(\mathbf{X}^e|\mathbf{Z}) \propto \frac{1}{|\mathbf{Z}|^{\frac{\theta^a}{2}+n+1}} \mathrm{e}^{-\frac{\theta^{\mathrm{a}}}{2}\,\mathrm{tr}(\mathbf{Z}^{-1}\mathbf{Z}^{\mathrm{a}})}, \tag{B.4}$$

where

$$\theta^a = \theta + N \qquad \text{and} \qquad \mathbf{Z}^a = \frac{\theta\bar{\mathbf{Z}} + N\mathbf{S}}{\theta + N}. \tag{B.5}$$

In the right-hand side of Eq.(B.4), we recognize again the inverse Wishart pdf (hence its conjugacy), see Eq.(A.2), with θ^a being the posterior sharpness parameter and $\mathbf{Z}^a$ being the posterior mean of $\mathbf{Z}$. Consequently, $\mathbf{Z}^a$ is the mean-square optimal point estimate of $\mathbf{Z}$ given both the prior and ensemble information. So, we have optimally assimilated the (conditionally Gaussian) ensemble data to update the (inverse Wishart) prior distribution of the random covariance matrix.

AppendixC. Integral w.r.t. a matrix

For a general $n \times n$-matrix $\mathbf{C}$, the integral $\int f(\mathbf{C})\,\mathrm{d}\mathbf{C}$ of a scalar function $f(\mathbf{C})$ over the space of all matrices with real entries is defined as follows. First, we *vectorize* $\mathbf{C}$, i.e. build the vector $\vec{\mathbf{C}}$ of length n^2 that comprises all entries of $\mathbf{C}$. Then, we simply identify $\int f(\mathbf{C})\,\mathrm{d}\mathbf{C}$ with $\int f(\mathbf{C})\,\mathrm{d}\vec{\mathbf{C}}$, that is, with the traditional multiple (Lebesgue or Riemann) integral over the Euclidean space of dimensionality n^2.

The integral w.r.t. a symmetric positive definite matrix is defined in a similar way. The difference from the general matrix case is that the vectorization here involves collecting in $\vec{\mathbf{C}}$ only algebraically independent matrix entries (e.g. the upper triangle of $\mathbf{C}$) and the multiple integral is over the set (the convex cone) of those $\vec{\mathbf{C}}$ that correspond to positive definite matrices.

AppendixD. List of main symbols

$()^a$	posterior (analysis) pdf (i.e. conditioned on past and current data) and its parameters
$()^f$	prior (forecast) pdf (i.e. conditioned on past data) and its parameters
$\widetilde{()}$	sub-posterior pdf (i.e. conditioned on past data and current ensemble data) and its parameters
$()^{fe}, ()^{ae}$	forecast ensemble / analysis ensemble
$()^{me}$, $()^{pe}$	model error ensemble / predictability ensemble
$\bar{\cdot}$	time mean value
$\langle\cdot\rangle$	average over L independent realizations of the truth / assimilation runs
$\mathbf{A}$	analysis error (posterior) covariance matrix
$\mathbf{B}$	forecast error (prior) covariance matrices
$\mathbf{F}$	forecast operator
$\mathbf{H}$	observation operator
i	ensemble member index
$\mathbf{K}$	Kalman gain matrix
k	time instance index
L	number of independent assimilation runs
$l(\mathbf{B}\|\mathbf{y})$	observation likelihood of the matrix $\mathbf{B}$
$\mathbf{m}^a$	posterior mean conditioned on $\mathbf{P}, \mathbf{Q}$
n	dimensionality of state space
N	ensemble size
p	probability density function (pdf)
$\mathbf{P}, \mathbf{Q}, \mathbf{R}$	predictability error / model error / observation error covariance matrix
$\mathbf{S}$	sample (ensemble) covariance matrix
V_k	$\mathsf{Var}\, x_k$
$\mathbf{x}$	state vector, "truth"
$\mathbf{x}^a$	posterior mean vector and its approximations (deterministic analysis)
$\mathbf{x}^f$	prior mean vector (identified in this study with the deterministic forecast)
$\mathbf{x}^{\cdot\cdot}(i), \mathbf{x}^e(i)$	ensemble member
$\mathbf{X}$	ensemble
$\mathbf{y}$	observation vector
$\mathbf{Y}$	observation and ensemble data combined
$\mathcal{IW}$	inverse Wishart distribution (parametrized according to AppendixA)
$\mathcal{N}(\mathbf{m}, \mathbf{B})$	Gaussian distribution with the mean $\mathbf{m}$ and (co)variance $\mathbf{B}$
$\boldsymbol{\varepsilon}$	model error (system noise) vector
$\boldsymbol{\eta}$	observation error vector

θ, ϕ, χ	sharpness parameters for the inverse Wishart pdfs
π	portion of time the process F_k is greater than 1 in modulus
σ	(time-specific) model error standard deviation
$\mathbb{E}$	expectation operator
rms, RMSE	root-mean-square value / error
$\mathsf{s.d.}$, Var	standard deviation / variance
tr	matrix trace
$\propto$	proportionality
$\sim$	has (corresponds to) the probability distribution
$1:k$	concatenation from the time instance 1 to the time instance k

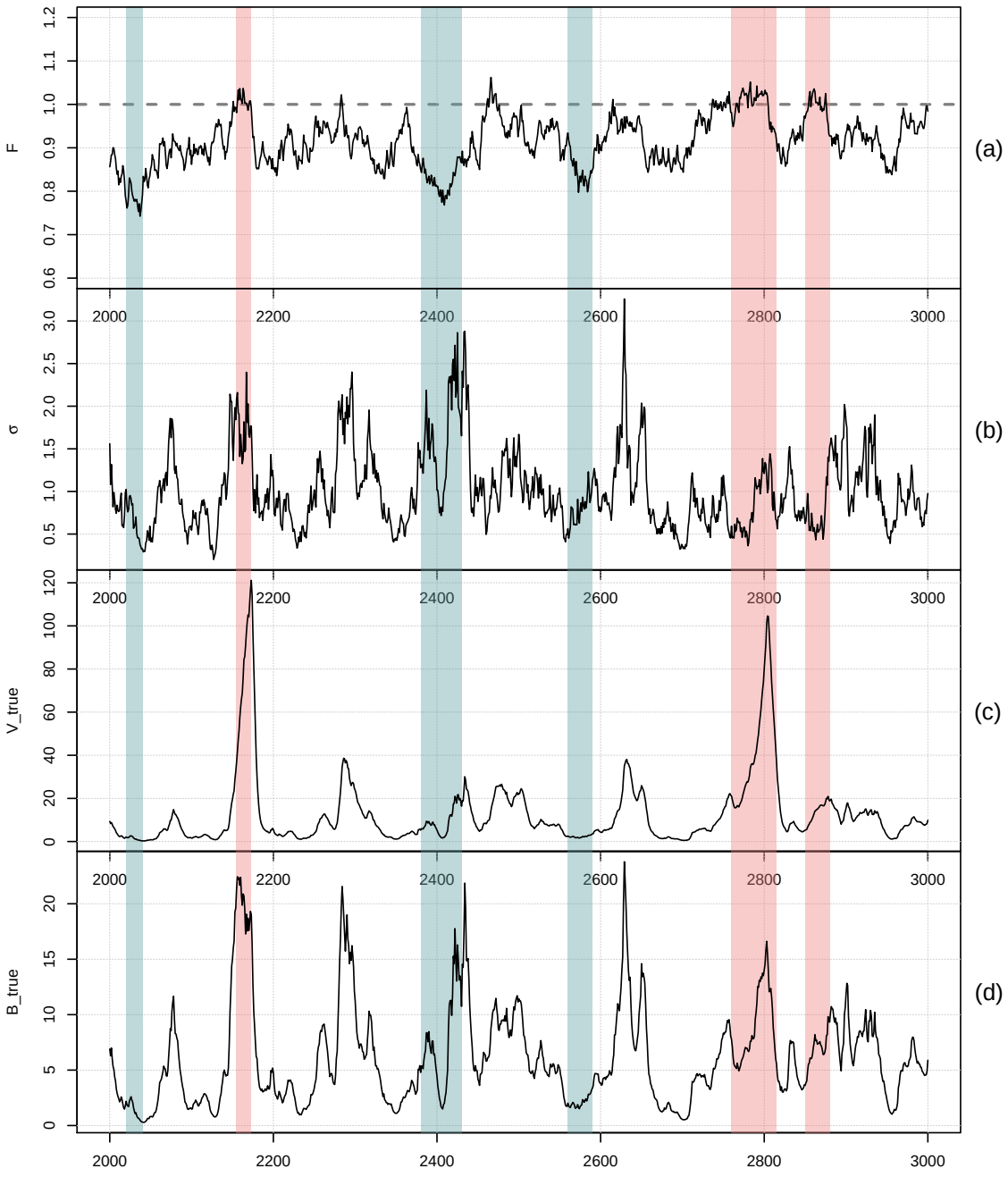

Figure 1: Typical time series of: (a) The forecast operator F_k, (b) The model error standard deviation σ_k, (c) The variance of the truth V_k, and (d) The true background error variance B_k for the HBEF.

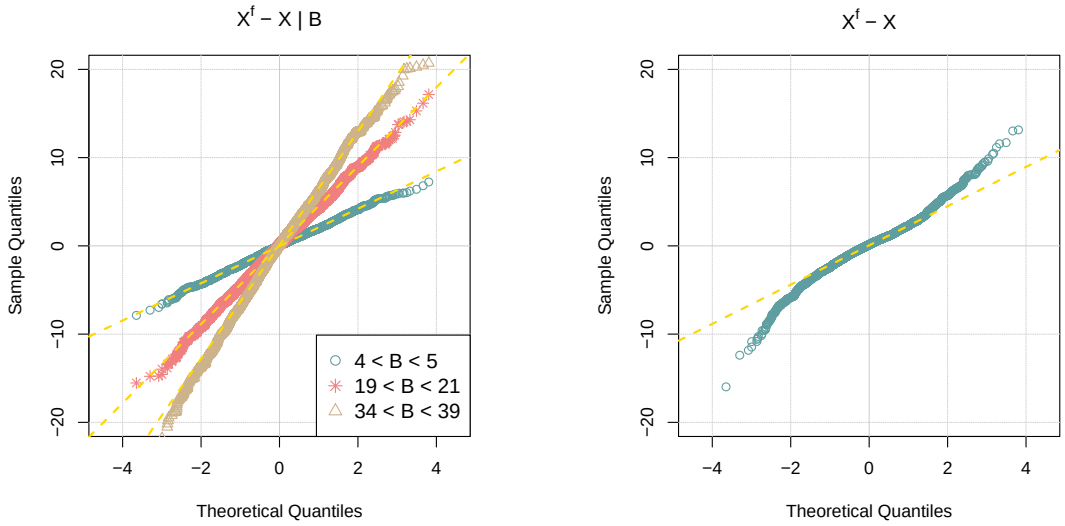

Figure 2: The q-q plots for the conditional pdf $p(x_k^f - x_k|B_k)$ (left) and the unconditional pdf $p(x_k^f - x_k)$ (right). In the left panel, the three curves correspond to the three intervals of B_k indicated in the legend.

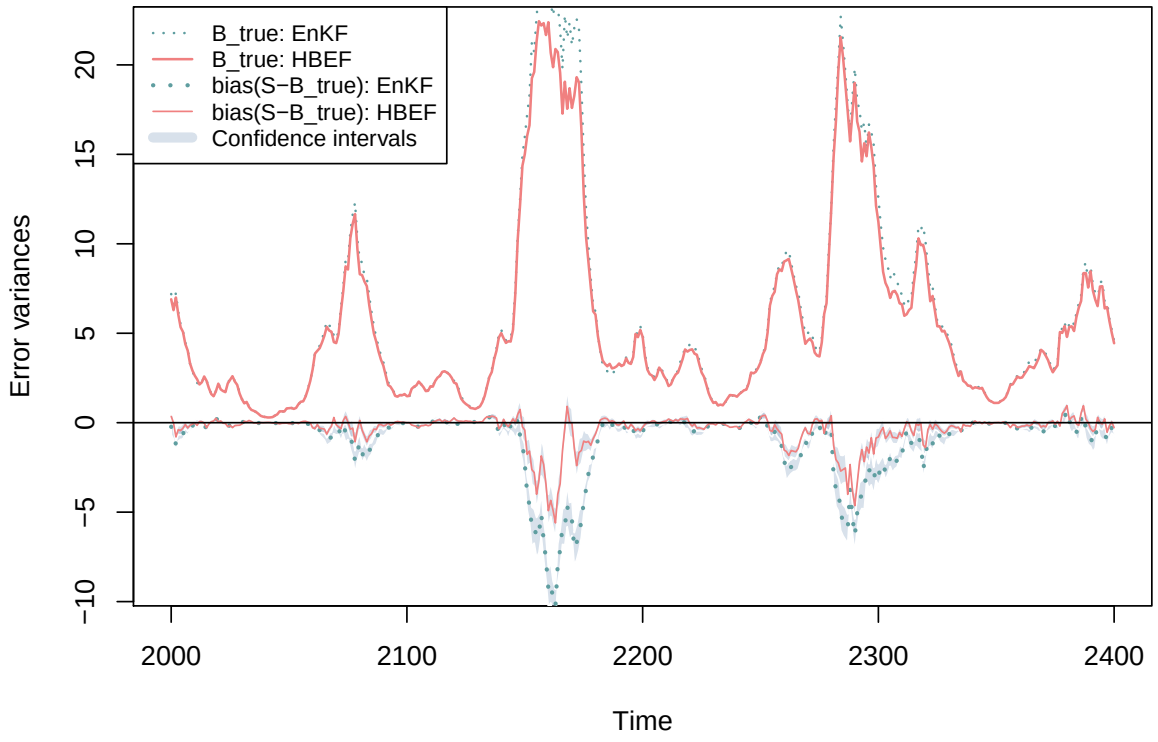

Figure 3: The two lower curves: biases in the forecast ensemble variances (with the 95% confidence intervals) for the EnKF and the HBEF. The two upper curves: the respective true error variances B_k.

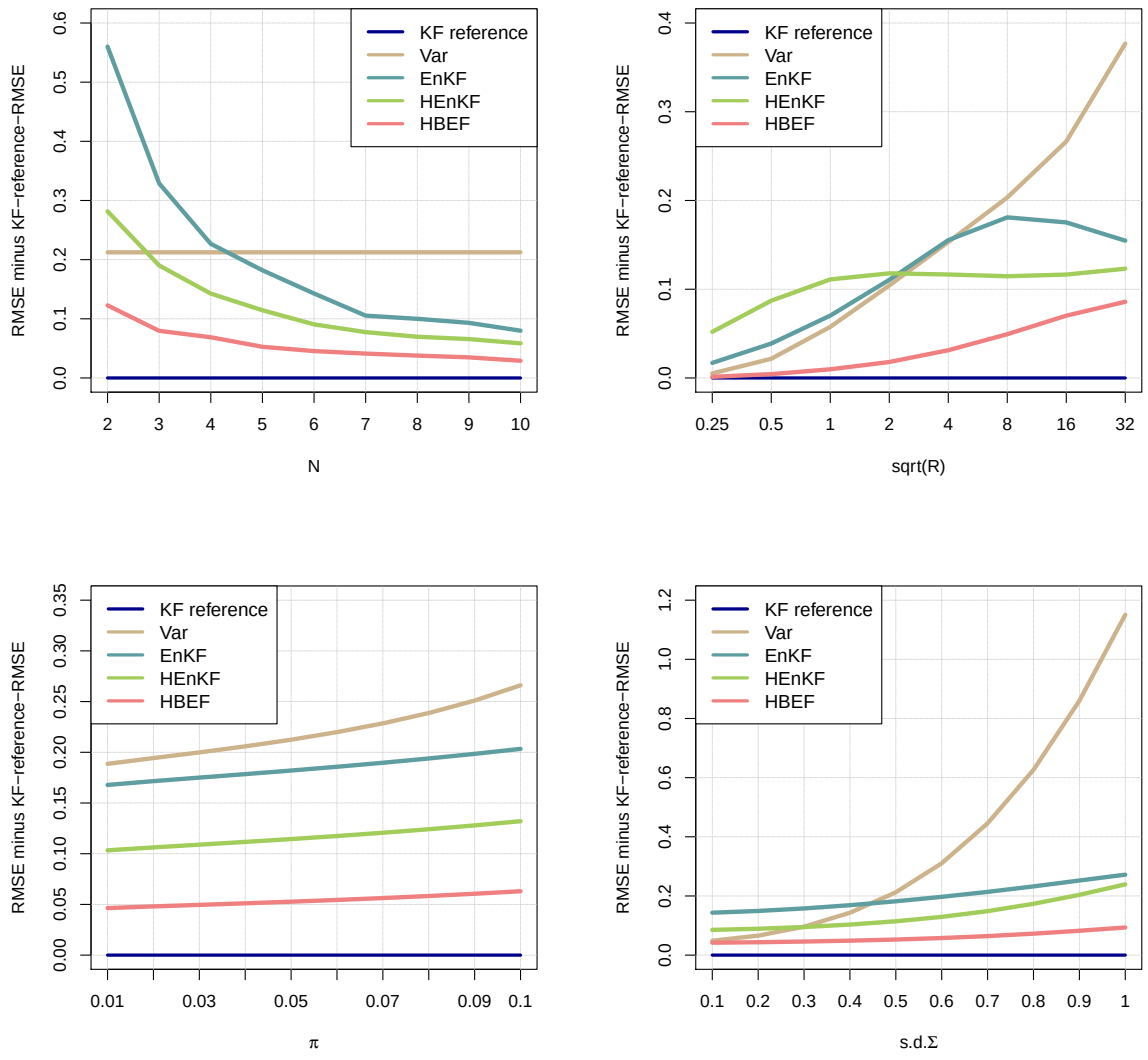

Figure 4: The filters' analysis RMSEs of the state (with the reference-KF analysis RMSE subtracted) as functions of the ensemble size N (top, left), the observation error standard deviation $\sqrt{R}$ (top, right), the degree of the system's intermittent instability π (bottom, left), the variability in the model error standard deviation $\mathsf{s.d.}\,\Sigma$ (bottom, right).

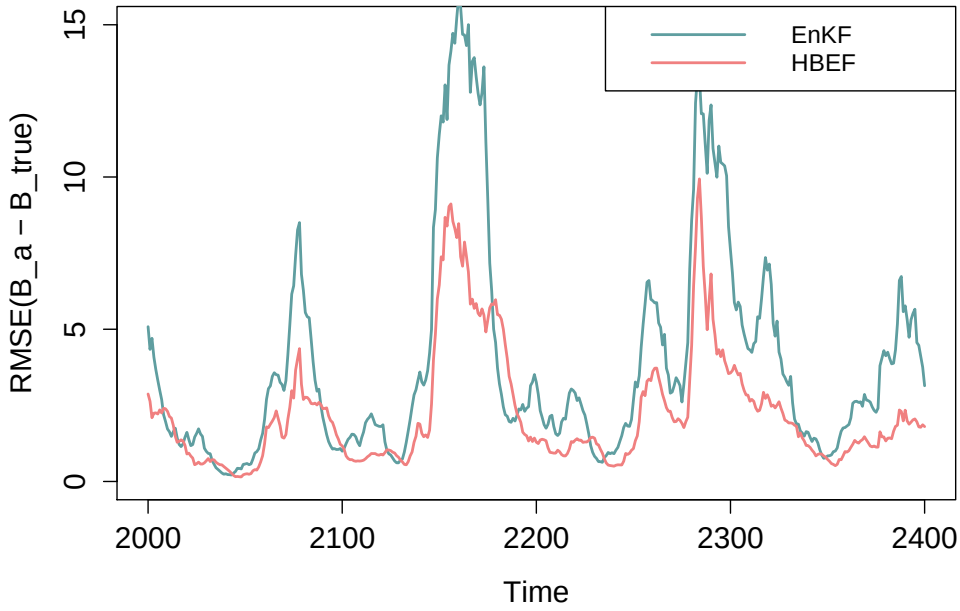

Figure 5: RMSEs in B_k^a produced by the EnKF and the HBEF.

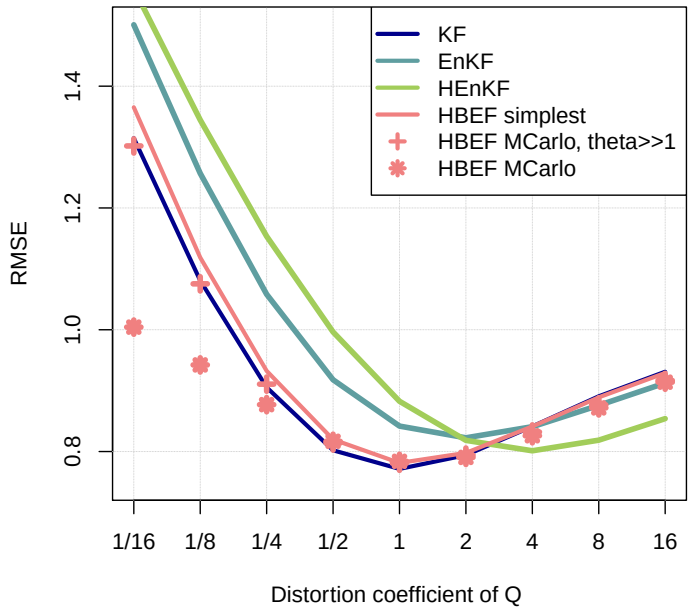

Figure 6: Analysis RMSEs of the state as functions of $q_{distort}$ for the filters with the distorted $Q = Q_{true} \cdot q_{distort}$.