Redesigning the ensemble Kalman filter
with a dedicated model of epistemic uncertainty

Chatchuea Kimchaiwong[1], Jeremie Houssineau[2], and Adam M. Johansen[3]

[1]Warwick Mathematics Institute, University of Warwick, UK.
[2]Division of Mathematical Sciences, Nanyang Technological University, Singapore.
[3]Department of Statistics, University of Warwick, UK.

Abstract

The problem of incorporating information from observations received serially in time is widespread in the field of uncertainty quantification. Within a probabilistic framework, such problems can be addressed using standard filtering techniques. However, in many real-world problems, some (or all) of the uncertainty is epistemic, arising from a lack of knowledge, and is difficult to model probabilistically. This paper introduces a *possibilistic* ensemble Kalman filter designed for this setting and characterizes some of its properties. Using possibility theory to describe epistemic uncertainty is appealing from a philosophical perspective, and it is easy to justify certain heuristics often employed in standard ensemble Kalman filters as principled approaches to capturing uncertainty within it. The possibilistic approach motivates a robust mechanism for characterizing uncertainty which shows good performance with small sample sizes, and can outperform standard ensemble Kalman filters at given sample size, even when dealing with genuinely aleatoric uncertainty.

Keywords: Bayesian inference; State-Space Model; Possibility Theory

1 Introduction

Dynamical real-world systems are typically represented as a hidden process in some state space, indirectly observed via a measurement process. The usual approach to characterizing the unknown state and uncertainty about it is to combine the observed information with given prior information in an approach known as filtering, or sometimes *data assimilation* [29]. However, this is non-trivial due not only to imprecise measurements, but also to model misspecification [31].

Perhaps the best-known data-assimilation algorithm is the Kalman filter (KF) [22], which is the optimal filter for the linear Gaussian state-space model, as detailed, e.g., in [31]. However, its application is limited by its high computational costs when dealing with large state spaces, and by the fact that it cannot be applied directly to nonlinear models [28]. Several modifications of the KF have been developed to extend its scope. The extended Kalman Filter can deal with a mild level of nonlinearity by approximating the model with a first order Taylor series expansion, as described, e.g., in [34]. The unscented Kalman Filter (UKF) approximates the distribution with Gaussian distribution using a set of sigma points and their weights, often giving better accuracy for highly nonlinear models [37]. Still, the accuracy of the UKF depends upon the tuning of certain algorithmic parameters which determine the locations of the sigma points. Moreover, the number of sigma points (and hence the computational cost and accuracy) is fixed. Another widely used algorithm is the ensemble Kalman filter (EnKF) [5, 35], which uses an ensemble to approximate the distributions of interest. The strengths of the EnKF are that it does not require the calculation of the variance (avoiding costly matrix inversions in high-dimensional problems) and can be used with a nonlinear model without approximating it with its linear tangent model [28] as one simply needs to be able to sample

from the transition kernel of the dynamic model—although this comes at the cost of further approximating a non-Gaussian distribution with a Gaussian one. The EnKF must be used carefully with high-dimensional problems as constraints on computing power typically lead to an ensemble size relatively small compared to the dimension of the state space, leading to problems such as spurious correlation and underestimated variance, affecting the algorithm's performance. Thus, the recent development in this field mainly focuses on reducing computational cost and increasing the forecast's accuracy.

Although there has been much recent development in the field of data assimilation, e.g., [1, 2, 23], the opportunities offered by alternative representations of the uncertainty have not been fully explored. Yet, there is a clear motivation for doing so: In many situations in which the EnKF is used, the main sources of uncertainty are i) the initial state of the system, ii) unknown deviations in the model, e.g., unknown forcing terms, and iii) the lack of data. These sources of uncertainty can be argued to be of an epistemic nature and, given their predominance, it is important to model them as faithfully as possible. This makes appealing the use of a dedicated framework for epistemic uncertainty. Many different approaches have been suggested in the literature to deal with epistemic uncertainty, such as fiducial inference [13, 16] and Dempster–Shafer theory [9]. While fiducial inference gives a probabilistic statement about a parameter without relying on a prior probability distribution, it still represents the information a posteriori via probability distributions. On the other hand, Dempster–Shafer theory leverages (fuzzy) set-valued probabilities to represent both aleatoric and epistemic uncertainty [27]. Yet, its formulation, based on sums over all subsets of the state space, restricts its application to discrete cases and, even then, does not scale easily to larger problems. Another approach, termed possibility theory [8, 11, 10], proposes to model exclusively epistemic uncertainty by changing the algebra underpinning probability theory, i.e., summation/integration is replaced by maximisation. Many probabilistic concepts and analyses can be adapted to possibility theory, for instance, [20] proves a Bernstein-von Mises theorem for a possibilistic version of Bayesian inference. In real systems, some uncertainty will be well modelled as aleatoric and some will of an epistemic nature; [17] shows that probability theory and possibility theory can be combined in this case, offering an extension of the probabilistic approach rather than an alternative. Possibility theory can be applied to complex problems, for instance [18] introduces an analogue to spatial point processes, and [19] shows that the Kalman filter can also be derived in this framework. These results highlight the potential of possibility theory for data assimilation, which we explore in this work via a possibilistic analogue of the EnKF. Related work includes [4] in which a robust version of the Kalman filter is proposed based on sets of probability distributions, with a motivation close to the one for this work.

Our contributions are as follows:

1. We introduce a new notion of Gaussian fitting for possibility theory, and contrast it against the standard moment-matching procedure. We highlight the potential of this approach for providing principle grounds for standard heuristics used in the EnKF.

2. We derive a possibilistic analogue of the EnKF, adapting the initialisation and predictions steps to the assumed epistemic uncertainty in the model, and showing how the update step of existing versions of the EnKF remain valid in this context.

3. We assess the performance of our method against two versions of the EnKF and against the UKF, considering both linear and nonlinear dynamics as well as fully and partially observed processes.

The remainder of this paper is organised as follows. Section 2 starts with the state estimation problem for the state-space model before giving the filtering techniques used to estimate such states under the probabilistic framework. Then, the details of the possibilistic framework developed to tackle epistemic uncertainty are given, with some concepts defined analogously to the probabilistic framework. Next, the possibilistic EnKF is introduced in Section 3 to estimate the system's state in the presence of epistemic uncertainty. After that, the performance of our algorithm is assessed on simulated data in a range of situations and compared against standard baselines in Section 4.

2 Background

We first briefly review the standard EnKF and the underlying filtering problem in Section 2.1 before introducing, in Section 2.2, the framework based on which the proposed method will be derived.

2.1 State estimation problem and filtering technique

The standard, probabilistic Gaussian state-space model describes the evolution of a hidden *state* process, $(X_k)_{k>0}$, and its associated observation process, $(Y_k)_{k>0}$, via

$$X_k = F_k(X_{k-1}) + \epsilon_k \qquad \text{where} \quad \epsilon_k \sim \mathrm{N}(0, U_k) \qquad (1\text{a})$$

$$Y_k = H_k(X_k) + \varepsilon_k \qquad \text{where} \quad \varepsilon_k \sim \mathrm{N}(0, V_k), \qquad (1\text{b})$$

with U_k and V_k positive definite matrices and with $\mathrm{N}(\mu, \Sigma)$ denoting a Gaussian distribution of mean μ and covariance Σ.

We consider the setting in which a stream of observations $y_1, y_2, \ldots$ arrives sequentially in time, and we want to obtain estimates of the hidden states as they arrive. Thus, in the Bayesian context, we characterise our knowledge of the state X_k at time step $k > 0$ given the realisations $y_{1:k} \doteq (y_1, y_2, \ldots, y_k)$ of the observations $Y_{1:k}$ via the filtering density $p(x_k|y_{1:k})$ which can be computed recursively, see, e.g., [31], as

$$p(x_k|y_{1:k-1}) = \int p(x_k|x_{k-1})p(x_{k-1}|y_{1:k-1})\mathrm{d}x_{k-1} \qquad (\text{prediction})$$

$$p(x_k|y_{1:k}) = \frac{p(y_k|x_k)p(x_k|y_{1:k-1})}{\int p(y_k|x_k)p(x_k|y_{1:k-1})\mathrm{d}x_k}. \qquad (\text{update})$$

The Kalman Filter (KF) provides a closed-form recursion for the prediction and update steps for linear Gaussian state-space models (see, e.g., [31]). However, many models in this world are nonlinear, meaning that the KF is not directly applicable and, unfortunately, there are few other settings in which analytic recursions are available. This motivates the development of numerous alternatives, including the UKF and the EnKF.

While the UKF can provide good performance for moderately nonlinear models, which depends upon careful choice of several algorithmic tuning parameters, and, as the number of sigma points is fixed, there is little scope for adjusting computational cost or accuracy. The EnKF algorithm takes a different approach; it can be viewed as an approximation to the KF which employs an ensemble of samples $\{X_k^i\}_{i=1}^N$ to approximate the parameters of the Gaussian distribution at time k. In particular, the ensemble's mean and variance are used to approximate those of the filtering distribution. The EnKF algorithm then proceeds recursively via the prediction and update steps. The prediction step mirrors that of the KF but uses an ensemble approximation, which means that F_k need not be linear as long as it can be evaluated pointwise as in Algorithm 1 [5, 35].

Algorithm 1 Prediction step of EnKF at time step k

Input: The posterior ensemble $\{\hat{X}_{k-1}^i\}_{i=1}^N$ at time step $k-1$.
Output: The predictive ensemble $\{X_k^i\}_{i=1}^N$ at time step k with mean μ_k and variance Σ_k.
1: **for** $i \in \{1, \ldots, N\}$ **do**
2:     $\epsilon_k^i \sim \mathrm{N}(0, U_k)$ \\Sample noise
3:     $X_k^i = F_k(\hat{X}_{k-1}^i) + \epsilon_k^i$ \\Predict i-th particle
4: **end for**
5: $\mu_k \leftarrow \frac{1}{N}\sum_{i=1}^N X_k^i$ \\Predictive mean
6: $\Sigma_k \leftarrow \frac{1}{N-1}\sum_{i=1}^N (X_k^i - \mu_k)(X_k^i - \mu_k)^\top,$ \\Predictive variance

---

The update step of the EnKF also approximates that of the KF: the mean of the updated distribution is a convex combination of that of the predictive distribution and the observation using the so-called *Kalman*

gain as a weight. However, using the traditional Kalman gain can lead to an underestimation in the posterior variance (as it essentially neglects the observation noise) and can lead to filter divergence [38]. A number of variants of the EnKF update step have been developed to address this problem, including both the stochastic EnKF (StEnKF) and the square root EnKF (SqrtEnKF) [24]. The latter updates the deviation of the ensemble by applying a matrix obtained from the square root of a certain matrix. Different variations of the SqrtEnKF exist, depending on how such a matrix is obtained, as summarised in [35]. In this work, we consider the version of the SqrtEnKF presented in [38], which is particularly suitable as a basis for introducing a possibilistic version of the EnKF. The update of the SqrtEnKF for a linear observation model is detailed in Algorithm 2, where the notation $A^{1/2}$ refers to the Cholesky factor of matrix A, and where we use the same notation for a linear function and for its matrix representation.

Algorithm 2 Update step of SqrtEnKF at time step k

Input: The predictive ensemble $\{X_k^i\}_{i=1}^N$ at time step k with mean μ_k and variance Σ_k.
Output: The posterior ensemble $\{\hat{X}_k^i\}_{i=1}^N$ at time step k.
1: $S_k \leftarrow H_k \Sigma_k H_k^\top + V_k$ \\Covariance of the innovation from the ensemble
2: $K_k \leftarrow \Sigma_k H_k^\top S_k^{-1}$ \\Standard Kalman gain
3: $\tilde{K}_k \leftarrow \Sigma_k H_k^\top \left(S_k^{1/2}\right)^{-\top} \left(S_k^{1/2} + V_k^{1/2}\right)^{-1}$ \\Adjusted Kalman gain
4: $\hat{\mu}_k \leftarrow \mu_k + K_k(Y_k - H_k \mu_k)$ \\Posterior mean
5: **for** $i \in \{1, \ldots, N\}$ **do**
6: $\quad \hat{e}_k^i \leftarrow (I_n - \tilde{K}_k H_k)(X_k^i - \mu_k)$ \\Posterior deviation
7: $\quad \hat{X}_k^i \leftarrow \hat{e}_k^i + \hat{\mu}_k$ \\Updated particle
8: **end for**

For the ensemble-based method, the accuracy of the state estimate depends highly on the ensemble size. The situation in which the ensemble is too small to adequately represent the system is termed undersampling and leads to four main problems which have been termed: inbreeding, spurious correlation, matrix singularity, and filter divergence [28]. Two widely used techniques to negate the undersampling problems are inflation and localisation [28]. The inflation technique increases the deviation between the sample and the predictive mean so that the predictive covariance from the ensemble is not underestimated and avoids the so-called inbreeding problem. On the other hand, the localisation techniques are implemented to eradicate the spurious correlation and singular matrix, examples being tapering using the Schur product to cut off the long-range correlation or domain localisation where assimilation is performed in a local domain, which is disjoint in a physical space [21]. Each technique needs to be fine-tuned in a problem-specific manner to achieve good accuracy.

We have focused here on approaches directly relevant to the methodology developed below in Section 3; many other methods have been devised to address nonlinearity and non-Gaussianity—see, e.g., [2, 12, 36].

2.2 Review of possibility theory

Possibility theory aims to directly capture epistemic uncertainty about a fixed but unknown element θ^* in a set Θ, by focusing on the *possibility* of an event related to θ^* rather than on its probability. A possibility of 1 corresponds to the absence of evidence against the event taking place *not* evidence that the event took place almost surely. If there is no evidence against an event, say E, then there cannot be any evidence against (E or E'), with E' any other event, so that the possibility of (E or E') is also equal to 1. This behaviour shows the notion of possibility does not give an additive measure of uncertainty, and the simplest operation that corresponds to the possibility of a union of events is the *maximum* of their individual possibilities.

The analogue of a probability density is a so-called possibility function, $f : \Theta \to [0, 1]$ which must have supremum 1. Just as probabilities are best described by measures, possibilities are naturally cast as outer measures, $\bar{P}$ and just as one may obtain a probability from a density one can obtain a possibility from a possibility function by taking its supremum, that is, for any $A \subseteq \Theta$, $\bar{P}(A) = \sup_{\theta \in A} f(\theta)$. If it holds that $f(\theta) = 1$ for some $\theta \in A$, then we will indeed have $\bar{P}(A) = \bar{P}(A \cup B) = 1$ for any $B \subseteq \Theta$, as required.

Formally, the set function $\bar{P}$ is an outer measure verifying $\bar{P}(\Theta) = 1$; we therefore refer to such set functions as *outer probability measures* (o.p.m.s). If $\bar{P}(A) = 1$ models the absence of opposing evidence, $\bar{P}(A) = 0$ has essentially the same interpretation as with probabilities, i.e., that $\theta^* \in A$ is impossible, which requires very strong evidence. In general, $\bar{P}(A)$ can be interpreted as an upper bound for *subjective* probabilities of the event $\theta^* \in A$, i.e., the maximum probability that we would be ready to assign to this event is $\bar{P}(A)$. Under this interpretation, the statement $\bar{P}(A) = 1$ is uninformative: we could assign any probability in the interval $[0, 1]$ to this event.

To formally define the quantities described above, we consider an analogue of a random variable, referred to as an *uncertain variable*, defined as a mapping $\boldsymbol{\theta} : \Omega \to \Theta$, and interpreted as follows: if the true outcome in Ω is ω then $\boldsymbol{\theta}(\omega)$ is the true value of the parameter. The set Ω plays a similar role to the elementary event space in probability theory, except that it is not equipped with any probabilistic structure; instead, it contains a true element ω^* and it holds by construction that $\boldsymbol{\theta}(\omega^*) = \theta^*$. An event can now be defined as a subset of Ω, for instance $\{\omega \in \Omega : \boldsymbol{\theta}(\omega) \in A\}$ for some $A \subseteq \Theta$. We will use the standard shortcut and write this event as $\boldsymbol{\theta} \in A$. An important difference between possibility theory and probability theory is that uncertain variables do not characterise possibility functions, instead possibility functions *describe* uncertain variables in a way that is not unique. For instance, if the available information is modelled by a possibility function f describing an uncertain variable $\boldsymbol{\theta}$, then any function g such that $g \geq f$, i.e. $g(\theta) \geq f(\theta)$ for any $\theta \in \Theta$, also describes $\boldsymbol{\theta}$. This is because g discarded some evidence about $\boldsymbol{\theta}$, that is, g is less informative than f. In particular, we can consider the possibility function equal to 1 everywhere, denoted $\mathbf{1}$, which upper bounds all other possibility functions. The function $\mathbf{1}$ models the total absence of information.

To gain information from data, Bayesian inference can be performed within possibility theory in a way that is very similar to the standard approach: If Y is a random variable with probability distribution $p(\cdot \mid \theta^*)$ belonging to a parameterised family of distributions[1] $\{p(\cdot \mid \theta) : \theta \in \Theta\}$, if we observe a realisation y of Y, and if the information available about θ^* a priori is modelled by the possibility function f, then the information available a posteriori can be modelled by the possibility function $f(\cdot \mid y)$ characterised by [17]

$$f(\theta \mid y) = \frac{p(y \mid \theta)f(\theta)}{\sup_{\theta' \in \Theta} p(y \mid \theta')f(\theta')}, \qquad \theta \in \Theta. \tag{2}$$

We borrow from the Bayesian nomenclature and simply refer to f and $f(\cdot \mid y)$ as the prior and the posterior. Since we can always start from the uninformative prior $f = \mathbf{1}$, it is easy to find posterior possibility functions by inserting different likelihoods; for instance, if $\Theta = \mathbb{R}^n$ and if the likelihood is a multivariate Gaussian distribution with mean θ and known variance Σ, i.e., $p(y \mid \theta) = \mathrm{N}(y; \theta, \Sigma)$, then

$$f(\theta \mid y) = \exp\left(-\frac{1}{2}(\theta - y)^\top \Sigma^{-1}(\theta - y) \right) \doteq \overline{\mathrm{N}}(\theta; y, \Sigma).$$

Such a Gaussian possibility function is a conjugate prior for the Gaussian likelihood as in the probabilistic case, and shares many of the properties of its probabilistic analogue. It can be advantageous to parameterise a Gaussian possibility function, say $\overline{\mathrm{N}}(\theta; \mu, \Sigma)$, by the precision matrix $\Lambda = \Sigma^{-1}$; indeed, the precision matrix does not need to be positive definite for the Gaussian possibility function to be well defined, with positive semi-definiteness being sufficient. In particular, this means that setting Λ to the 0 matrix is possible, with the Gaussian possibility function being equal to $\mathbf{1}$ in this case. A simple way to quantify the amount of epistemic uncertainty in a possibility function is to consider the integral $\int f(\theta)\mathrm{d}\theta$ as in [6], when defined. This notion of uncertainty is consistent with the partial order on possibility functions: if g is less informative than f then the integral of g will obviously be larger then that of f. In particular, the uncertainty of a Gaussian possibility function is $\sqrt{|2\pi\Sigma|}$, with $|\cdot|$ denoting the determinant; a quantity often used to quantify how informative a given Gaussian distribution is.

If $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ are two uncertain variables, respectively on sets Θ and Ψ, jointly described by a possibility function f on $\Theta \times \Psi$, then Bayes' rule can also be expressed via a more standard notion of conditioning as

---

[1]Although $p(\cdot \mid \theta)$ is not formally a conditional probability distribution, it is useful to slightly abuse notations and write it as such.

[8]

$$f(\theta \mid \psi) = \frac{f(\theta, \psi)}{\sup_{\theta' \in \Theta} f(\theta', \psi)},$$

where $\sup_{\theta' \in \Theta} f(\theta', \psi)$ is the marginal possibility function describing $\boldsymbol{\psi}$. The conditional possibility function $f(\theta \mid \psi)$ allows independence to be defined simply: if $f(\theta \mid \psi)$ is equal to the marginal $f(\theta) = \sup_{\psi' \in \Psi} f(\theta, \psi')$ for any $\theta \in \Theta$, then $\boldsymbol{\theta}$ is said to be independent from $\boldsymbol{\psi}$ under f. As with most concepts in possibility theory, the notion of independence depends on the choice of possibility function; if $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ are not independent under f, we could find another possibility function g such that $g \geq f$ and such that $\boldsymbol{\theta}$ is independent of $\boldsymbol{\psi}$ under g. This will be key later on for simplifying high-dimensional Gaussian possibility functions in a principled way. When there is no ambiguity about the underlying possibility function, we will simply say that $\boldsymbol{\theta}$ is independent of $\boldsymbol{\psi}$.

Although the parameters μ and Σ of the Gaussian possibility functions are reminiscent of the probabilistic notions of expected value and variance, these notions have to be redefined in the context of possibility theory. Asymptotic considerations [20] lead to a notion of expected value $\mathbb{E}^*(\cdot)$ based on the mode, that is $\mathbb{E}^*(\boldsymbol{\theta}) = \arg\max_{\theta \in \Theta} f(\theta)$, for any uncertain variable $\boldsymbol{\theta}$ described by f, and to a local notion of variance based on the curvature of f at $\mathbb{E}^*(\boldsymbol{\theta})$, assuming that the latter is a singleton, an assumption that we will make throughout this work. These quantities correspond to the approximation often referred to as Laplace/Gaussian approximation, and make sense in the context of epistemic uncertainty: without a notion of variability, we simply look at our best guess $\mathbb{E}^*(\boldsymbol{\theta})$ for the unknown θ^* and at how confident we are in that guess, i.e., how fast the possibility decreases around it. Although the expected value and variance associated with the Gaussian possibility function $\overline{N}(\mu, \Sigma)$ are indeed μ and Σ, they will differ in general from their probabilistic counterparts.

In order to make inference for dynamical systems simpler, we introduce a Markovian structure as follows: We consider a sequence of uncertain variables $\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots$ in a set X and assume that $\boldsymbol{x}_k$ is described by a possibility function f_k for some given $k \geq 0$. Following the standard approach, we assume that $\boldsymbol{x}_{k+1}$ is conditionally independent of $\boldsymbol{x}_{k-\delta}$ given $\boldsymbol{x}_k$, for any $\delta > 0$. The available information about $\boldsymbol{x}_{k+1}$ given a value x_k of $\boldsymbol{x}_k$ can then be modelled by a conditional possibility function $f_{k+1|k}(\cdot \mid x_k)$, and prediction can be performed via

$$f_{k+1}(x_{k+1}) = \sup_{x_k \in \mathsf{X}} f_{k+1|k}(x_{k+1} \mid x_k) f_k(x_k), \qquad x_{k+1} \in \mathsf{X}.$$

We will focus in particular on the case where an analogue of (1a) holds, that is

$$\boldsymbol{x}_k = F_k(\boldsymbol{x}_{k-1}) + \boldsymbol{u}_k,$$

with $\boldsymbol{u}_k$ an uncertain variable described by $\overline{N}(0, U_k)$. In this case, the term $\boldsymbol{u}_k$ cannot be interpreted as noise and models instead deviations between the model and the true dynamics: the true states x_k^* and x_{k-1}^* are related via $x_k^* = \tilde{F}_k(x_{k-1}^*)$ with $\tilde{F}_k$ potentially different from F_k; the true value u_k^* of interest is therefore equal to the difference $\tilde{F}_k(x_{k-1}^*) - F_k(x_{k-1}^*)$ between the model and the true dynamics.

In the case where the prediction is deterministic, that is $\boldsymbol{x}_{k+1} = F_k(\boldsymbol{x}_k)$ for some possibly non-linear mapping F_k, we can use the change of variable formula [3] to characterise the possibility function f_{k+1} as

$$f_{k+1}(x_{k+1}) = \sup \left\{ f_k(x_k) : x_k \in F_k^{-1}(x_{k+1}) \right\}, \tag{3}$$

where $F_k^{-1}(x_{k+1})$ is the (possibly set-valued) pre-image of x_{k+1} via F_k, and $\sup \emptyset = 0$ by convention. A result that is specific to possibility theory is that the expected values of $\boldsymbol{x}_{k+1}$ and $\boldsymbol{x}_k$ are related via $\mathbb{E}^*(\boldsymbol{x}_{k+1}) = F_k(\mathbb{E}^*(\boldsymbol{x}_k))$ without assumptions on F_k. This result will be key in our approach since it allows to compute the expected value at time step $k + 1$ with a single application of F_k, rather than averaging ensembles as in the EnKF. This result does not hold for the mode of probability densities because the corresponding change of variable formula include the Jacobian of the transformation, which shifts the mode in non-trivial ways.

6

3 Possibilistic EnKF

It is has been shown in [19] that an analogue of the Kalman filter can be derived in the context of possibility theory, and that the corresponding expected value and variance are the same as in the probabilistic case. However, as with the probabilistic KF, its applicability is limited to linear models—so it is natural to develop extensions that accommodate nonlinearity. Since the probabilistic EnKF is flexible in computational cost and can be effectively applied to nonlinear models, we explore the use of EnKF-like ideas in the possibilistic setting. Here, we present a novel *possibilistic EnKF* (p-EnKF) and show that analogues of inflation and localisation arise naturally in the context of possibility theory rather than being imposed as heuristics to improve performance as in the standard setting.

For the p-EnKF, we use an ensemble of weighted particles to characterise the possibility function and follow a similar path to the standard EnKF by assuming that the underlying possibility function is Gaussian in order to proceed with a KF-like update. For this purpose, we need to define two operations: 1) how to approximate a given possibility function by weighted particles? This will be necessary for initialisation at time step $k = 0$, and 2) how to define a Gaussian possibility function based on weighted particles? This will be necessary to carry out the Kalman-like update. We start by answering the latter question in Section 3.2, before moving to the former in Section 3.3. Based on this, we detail the prediction and update mechanism of the p-EnKF in Sections 3.4 and 3.5 respectively, and consider some extensions in Section 3.6.

3.1 Ensemble approximation

We consider the problem of defining a set $\{(w_i, x_i)\}_{i=1}^N$ of N weighted particles in $\mathbb{R}^n$, which will allow the approximate solution of optimisation problems of the form $\max_{x \in \mathbb{R}^n} \varphi(x) f(x)$, for some bounded function φ on $\mathbb{R}^n$ and for a given possibility function on $\mathbb{R}^n$. Although our approach could be easily formulated on more general spaces than $\mathbb{R}^n$, Euclidean spaces are sufficient for our purpose in this work and help to simplify the presentation. As is common in Monte Carlo approaches, we do not want the particles or weights to depend on φ, so as to allow us to solve this problem for many different such functions using a single sample, we therefore focus on directly approximating f: placing particles in locations which allow a good characterization of f allows us to approximate the optimization for a broad class of regular φ. Following the principles of Monte Carlo optimization [30, Chapter 5], and assuming that $\int f(x) \mathrm{d}x < +\infty$, we can sample from the probability distribution proportional to f to obtain the particles $\{x_i\}_{i=1}^N$ and then weight these particles with $w_i = f(x_i)$, for any $i \in \{1, \ldots, N\}$. We then obtain the approximation

$$\max_{x \in \mathbb{R}^n} \varphi(x) f(x) \approx \max_{i \in \{1, \ldots, N\}} w_i \varphi(x_i).$$

This approximation can be proved to converge when $N \to \infty$ under mild regularity conditions on φ as long as the supports of φ and f are not disjoint. This is a simple default choice for the methods developed below, although more sophisticated approaches are possible.

Remark 1. One initially appealing idea is to consider a form of maximum entropy principle by identifying the constraints we want to impose on the probability distributions from which we may wish to sample, and to pick the distribution of maximum entropy subject to satisfying these constraints. This is a standard argument for specifying the "least informative" distribution with particular properties, e.g., given a distribution when the mean and variance the maximum entropy distribution is the Gaussian with these moments. In the considered context, the constraints comes from the interpretation of o.p.m.s as upper bounds for probability distributions. Specifically, for an o.p.m. $\bar{P}$, we could draw an ensemble from the probability density p satisfying the following criteria: p has the maximum entropy amongst those distributions satisfying $\int_A p(x) dx \leq \bar{P}(A)$ for any $A \subseteq \mathbb{R}^n$. The distribution p would typically be more diffuse than the one proportional to f, which can be beneficial for optimizing functions φ taking large values in the "tails" of f. However, obtaining such a distribution is in general highly non-trivial beyond discrete or univariate problem.

As there are no constraints on the way in which particle locations, $\{x_i\}_{i=1}^N$ are selected, deterministic schemes such as quasi Monte Carlo [26, 15] could also be considered. In Section 4, we will consider using

the same approach as in the UKF to define particle locations, which will prove to be beneficial in some situations.

In the situations that we will consider, it will always be the case that f is maximised at a single known element $x^* \in \mathbb{R}^n$. Therefore, it makes sense to add one particle $x_0 = x^*$ with weight $w_0 = 1$. This will prove beneficial when fitting a Gaussian possibility function to the ensemble $\{(w_i, x_i)\}_{i=0}^N$, as detailed in the next section.

3.2 Best-fitting Gaussian possibility function

We consider the situation in which the epistemic uncertainty is captured by a set $\{(w_i, x_i)\}_{i=0}^N$ of $N+1$ weighted particles in $\mathbb{R}^n$, with $w_i = 1$ if and only if $i = 0$. We denote by $\tilde{f}$ the "empirical" possibility function defined based on the ensemble as $\tilde{f}(x) = \max_{i \in \{0,\dots,N\}} w_i \mathbf{1}_{x_i}(x)$, with $\mathbf{1}_{x_i}$ the indicator of the point x_i. The standard approach would be to simply compute the (weighted) mean and variance of the ensemble, but these notions do not apply directly to possibility functions, and the curvature-based possibilistic notion of variance is not defined for an empirical possibility function like $\tilde{f}$. Instead, we aim to fit a Gaussian possibility function $\overline{\mathrm{N}}(\mu, \Sigma)$ to this (weighted) ensemble and the considered variance will simply be the second parameter Σ of this fitted Gaussian.

To fit the Gaussian $\overline{\mathrm{N}}(\mu, \Sigma)$ to the ensemble, we need a notion of best fit. Based on the partial order between possibility functions described in Section 2.2, we can easily ensure that $\overline{\mathrm{N}}(\mu, \Sigma)$ does not introduce artificial information—at least at locations $\{x_i\}_{i=0}^N$ – by requiring that $\overline{\mathrm{N}}(\mu, \Sigma) \geq \tilde{f}$. Since $\tilde{f}(x) = 0$ when $x \notin \{x_i\}_{i=0}^N$, we can simplify this condition to: $\overline{\mathrm{N}}(x_i; \mu, \Sigma) \geq w_i$ for all $i \in \{0, \dots, N\}$. The inequality $\overline{\mathrm{N}}(\mu, \Sigma) \geq \tilde{f}$ forces the expected values of $\overline{\mathrm{N}}(\mu, \Sigma)$ and $\tilde{f}$ to coincide, from which we can deduce that $\mu = x_0$.

For the variance, we aim to minimise the uncertainty in the possibility function $\overline{\mathrm{N}}(\mu, \Sigma)$, i.e., minimising $\int \overline{\mathrm{N}}(x; \mu, \Sigma) \mathrm{d}x \propto \sqrt{|\Sigma|}$, which is equivalent to maximising $\log |\Lambda|$ with $\Lambda = \Sigma^{-1}$ the precision matrix, thanks to the properties of the determinant. The precision matrix is then most naturally defined as the one solving the constrained optimisation problem

$$\max_{\Lambda \in \mathbf{S}_+^d} \log |\Lambda| \qquad \text{subject to} \quad \overline{\mathrm{N}}(x_i; \mu, \Sigma) \geq w_i, \quad 1 \leq i \leq N, \tag{4}$$

where $\mathbf{S}_+^d$ is the cone of positive semi-definite $d \times d$ matrices. Since $\mu = x_0$, the corresponding constraint $\overline{\mathrm{N}}(x_0; \mu, \Sigma) \geq w_0$ is automatically satisfied and we only need to ensure our Gaussian possibility function upper bounds the ensemble at the other x_i's. Using the invariance of the trace to cyclic permutations allows these constraints to be rewritten as

$$(x_i - \mu)^\top \Lambda (x_i - \mu) = \mathrm{Tr}\,(C_i \Lambda) \leq -2 \log w_i, \qquad 1 \leq i \leq N, \tag{5}$$

where $C_i = (x_i - \mu)(x_i - \mu)^\top$; this is more convenient for numerical optimization because $\mathrm{Tr}(C_i \Lambda)$ is linear in Λ.

Remark 2. In the one-dimensional ($n = 1$) case, this optimisation problem can be easily solved for any $N \geq 1$: the ith constraint in (5) can be expressed directly for the (scalar) variance as $\sigma^2 \doteq \Sigma \leq -2 \log w_i / (x^i - \mu)^2$, and (5) reduces to a single constraint:

$$\sigma^2 \leq \min_{i \in \{1, \dots, N\}} \frac{-2 \log w_i}{(x_i - \mu)^2}. \tag{6}$$

In particular, setting σ^2 to the right hand side of this inequality will maximise the precision, hence solving the optimisation problem (4). In this case, if it happens that $w_i = f(x_i)$ with f a Gaussian possibility function, then the associated variance will be recovered exactly, even with $N = 1$. Although this result does not generalise easily to higher dimensions, it highlights the potential of this method for recovering the variance from an ensemble, as will be studied later in this section.

We will denote by $\Lambda^*\big(\{(w_i, x_i)\}_i\big)$ the solution to (4), omitting the limits on i for concision. In addition to providing an estimate for the variance of the best-fitting Gaussian, this approach provides a measure

of how far from Gaussian the ensemble is. Indeed, a notion of distance can be defined based on the gaps $-\operatorname{Tr}(C_i \Lambda) - 2 \log w_i$, $i \in \{1, \ldots, N\}$.

We will be interested in the relationship between the best Gaussian fit for a given ensemble and the one for a linear and invertible transformation of that ensemble, which motivates the following proposition.

Proposition 1. *Let $\{(w_i, x_i)\}_i$ be an ensemble such that $w_i = 1$ if and only if $i = 0$, and let M be an invertible linear map on $\mathbb{R}^n$, it holds that*
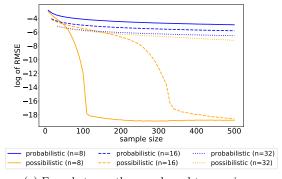
$$\Lambda^*\big(\{(w_i, x_i)\}_i\big) = M^\top \Lambda^*\big(\{(w_i, Mx_i)\}_i\big)M.$$
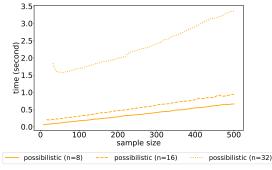
Proof. The constraints (5) for the ensemble (w_i, Mx_i) can be rewritten as

$$\forall i \in \{1, \ldots, N\}: \quad (Mx_i - Mx_0)^\top \Lambda (Mx_i - Mx_0) = (x_i - x_0)^\top M^\top \Lambda M (x_i - x_0) \leq -2 \log w_i.$$

Changing the variable of the optimisation problem from Λ to $\tilde{\Lambda} = M^\top \Lambda M$ changes the objective function to $\log |M^{-\top} \tilde{\Lambda} M^{-1}| = \log |\tilde{\Lambda}| + \text{constant}$. Therefore, the two optimisation problems are equivalent, and their solutions are related as stated. □

From a practical viewpoint, the computational cost for calculating the ensemble's variance via (4) is greater than that of the probabilistic framework due to the optimisation problem. Yet, Remark 2 hints at a potential for an efficient recovery of the true variance in the Gaussian case. This aspect is investigated in Figure 1a which displays the average root mean squared error (RMSE) between the true variance and the ensemble's variance based on different sample sizes and for $n \in \{8, 16, 32\}$, averaged over 1000 repeats. The true $n \times n$ covariance matrix Σ is drawn from the inverse Wishart distribution with n^2 degrees of freedom and scale matrix nI_n, with I_n the identity matrix of dimension $n \times n$, making its reconstruction appropriately challenging. Here, we use a sample $x_i \sim \mathrm{N}(0_n, \Sigma)$ for $i = 1, \ldots, N$, with 0_n the zero vector of size n, to both estimate the probabilistic variance and to be used as the ensemble in (4) with weights $w_i = \overline{\mathrm{N}}(x_i; 0_n, \Sigma)$. The performance of each case is investigated from the minimum required sample size $N = n$ up to $N = 500$. Figure 1a shows that the covariance matrix obtained from (4) indeed converges to the true underlying variance faster than the probabilistic one for small dimensions. In particular, the associated RMSE appears to drop significantly around $N = n^2$.

(a) Error between the sample and true variance

(b) Computational time of sample variance

Figure 1: Analysis of the proposed procedure for Gaussian fitting when the underlying possibility function / probability distribution is Gaussian.

The computational cost of the probabilistic approach is very small for the considered range of sample sizes and remains around $0.1\,\mathrm{ms}$ even for $n = 32$ in our experiments, which is likely due to other operations dominating the computational cost in the considered settings. This is in contrast to the computational time of the proposed method, shown in Figure 1b, which has a clear linear trend and is orders of magnitude larger than in the probabilistic case.

One advantage of defining an ensemble's variance via (4) is that the knowledge of dependency—or lack thereof—between components of the states $x \in \mathbb{R}^n$ can be more easily integrated. For instance, if we want to impose conditional independence between some components, i.e., to impose that $\Lambda_{i,j} = 0$ for any indices i and $j \neq i$ in a given set I, then we can add a set of constraints in the optimisation problem (4). Adding constraints will reduce the precision, i.e., it will yield a Gaussian possibility function that is less informative than the less-constrained problem. This behaviour can be interpreted as follows: in the possibilistic framework, conditional independence can be obtained by sacrificing some information. This is a well-understood trade-off in data assimilation, where forcing components to be uncorrelated, a method known as *localisation*, is usually accompanied by an increase of the variance of the remaining components, a process known as *inflation*. The main difference between the standard inflation and the proposed approach is that inflation usually comes with additional parameters that need to be fine-tuned for different situations, whereas the required amount of precision loss is automatically determined via the optimisation problem (4) when adding the constraints on conditional independence, with no additional tuning parameters.

Apart from defining the ensemble's expected value and variance, three steps of the algorithm need to be developed: initialisation, prediction, and update. These will be analogous to those of the probabilistic EnKF, but significant changes are required.

3.3   Initialisation

We consider a sequence of uncertain variables $\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots$, with $\boldsymbol{x}_k$ representing the state of the system at time step k. We assume that there is prior knowledge about $\boldsymbol{x}_0$, encoded into a possibility function $\overline{\mathrm{N}}(\mu_0, \Sigma_0)$. This possibility function is approximated by an ensemble $\{(w^i, \hat{x}_0^i)\}_{i=0}^N$ of $N + 1$ weighted particles, defined as in Section 3.1. The time index, 0 in this case, is omitted for the weights as these will in fact remain constant throughout the algorithm, which relies exclusively on transports of the associated particles. The initialisation step is detailed in Algorithm 3.

Algorithm 3 Initial step of the p-EnKF

Input: The prior Gaussian possibility function $\overline{\mathrm{N}}(\mu_0, \Sigma_0)$
Output: The initial ensemble $\{(w^i, \hat{x}_0^i)\}_{i=0}^N$
 1: **for** $i \in \{1, \ldots, N\}$ **do**
 2:    $\hat{x}_0^i \sim \mathrm{N}(\mu_0, \Sigma_0)$ \\Draw a particle
 3:    $w^i \leftarrow \overline{\mathrm{N}}(\hat{x}_0^i; \mu_0, \Sigma_0)$ \\Calculate the corresponding weight
 4: **end for**
 5: $\hat{x}_0^0 \leftarrow \mu_0$ \\Add a particle deterministically at the mode
 6: $w^0 \leftarrow 1$

3.4   Prediction step

In the probabilistic framework, the standard way of obtaining the predictive ensemble at time step k is to i) apply the transition model F_k to each particle, and ii) add a realisation of the transition noise. The first part of this process can be used as is, with some advantageous properties: a key result is the fact that $\mathbb{E}^*(F_k(\boldsymbol{x}_{k-1})) = F_k(\mathbb{E}^*(\boldsymbol{x}_{k-1}))$ with no assumption on either F_k or $\boldsymbol{x}_{k-1}$. This allows to perform the prediction without recomputing the expected value, therefore stabilising the estimation through non-linear dynamics. In practice, this means that the particle with index 0 will always correspond to the expected value, hence its special treatment. In practice, given the ensemble $\{(w^i, \hat{x}_{k-1}^i)\}_{i=0}^N$ at time step $k-1$, we can simply compute the image $\tilde{x}_k^i = F_k(\hat{x}_{k-1}^i)$ of each particle to capture the information at time k in the absence of perturbations.

However, the second part of the standard approach is not appropriate in the possibilistic setting since we aim to model epistemic uncertainty which, in this context, corresponds to deviations between the model and the actual dynamics rather than real perturbations. Deterministic methods can however be adapted and we

consider using transport maps to move the points of the ensemble as in [32]. To construct such a map for our setting, we first characterize linear transformations of uncertain variables as follows.

Proposition 2. *If $\boldsymbol{x}$ is an uncertain variable on $\mathbb{R}^n$ described by $\overline{\mathrm{N}}(\mu, \Sigma)$ and $\boldsymbol{z} = A\boldsymbol{x} + b$ where A is an $n \times n$ invertible matrix and b is a constant vector of length n, then $\boldsymbol{z}$ is described by $\overline{\mathrm{N}}(A\mu + b, A\Sigma A^\top)$.*

Proof. Let $f_{\boldsymbol{x}}$ denote the Gaussian possibility function $\overline{\mathrm{N}}(\mu, \Sigma)$. Applying (3), the possibility function of the uncertain variable $\boldsymbol{z}$ is

$$
\begin{aligned}
f_{\boldsymbol{z}}(z) &= \sup\{f_{\boldsymbol{x}}(x) : x \in \mathbb{R}^n, z = Ax + b\} \\
&= \sup\{f_{\boldsymbol{x}}(x) : x = A^{-1}(z - b)\} \\
&= f_{\boldsymbol{x}}(A^{-1}(z - b)) \\
&= \overline{\mathrm{N}}\left(A^{-1}(z - b); \mu, \Sigma\right) = \overline{\mathrm{N}}\left(z; A\mu + b, A\Sigma A^\top\right).
\end{aligned}
$$

$\square$

According to Proposition 2, Gaussianity is preserved under linear and invertible transformations in the possibilistic framework. Furthermore, it is straightforward to obtain a map between two uncertain variables in $\mathbb{R}^n$ described by Gaussian possibility functions:

Proposition 3 (Mapping between Gaussian possibility functions). *If $\boldsymbol{x}$ and $\boldsymbol{z}$ are two uncertain variables in $\mathbb{R}^n$ described respectively by $\overline{\mathrm{N}}(\mu, \Sigma)$ and $\overline{\mathrm{N}}(\tilde{\mu}, \tilde{\Sigma})$, then there exists a map M such that $\boldsymbol{z} = M(\boldsymbol{x})$, which is characterised by*

$$
M(x) = \tilde{\mu} + T(x - \mu), \qquad x \in \mathbb{R}^n
$$

where $T = \tilde{\Sigma}^{1/2}\left(\Sigma^{1/2}\right)^{-1}$, with the notation $A^{1/2}$ referring to the Cholesky factor of a matrix A.

Proof. First, we rewrite the mapping as follows

$$
M(\boldsymbol{x}) = \tilde{\mu} + T(\boldsymbol{x} - \mu) = T\boldsymbol{x} + (\tilde{\mu} - T\mu)
$$

As this is a linear and invertible transformation of $\boldsymbol{x}$, Proposition 2 guarantees that $\boldsymbol{z} = M(\boldsymbol{x})$ is also described by a Gaussian possibility function, defined as

$$
\begin{aligned}
f_{\boldsymbol{z}}(z) &= \overline{\mathrm{N}}\left(T\mu + (\tilde{\mu} - T\mu), T\Sigma T^\top\right) \\
&= \overline{\mathrm{N}}\left(\tilde{\mu}, \tilde{\Sigma}^{1/2}\left(\Sigma^{1/2}\right)^{-1}\left[\Sigma^{1/2}\left(\Sigma^{1/2}\right)^\top\right]\left(\Sigma^{1/2}\right)^{-\top}\left(\tilde{\Sigma}^{1/2}\right)^\top\right) = \overline{\mathrm{N}}\left(\tilde{\mu}, \tilde{\Sigma}\right).
\end{aligned}
$$

$\square$

Such a transport map can be used in the prediction step of the p-EnKF to add uncertainty from the transition model to the ensemble using deterministic mapping as follows: We fit a Gaussian possibility function $\overline{\mathrm{N}}(\mu_k, \tilde{\Sigma}_k)$ to the ensemble $\{(w^i, \tilde{x}_k^i)\}_{i=0}^N$, then, from the prediction of the possibilistic Kalman filter [19], we know that an additive Gaussian uncertainty of the form $\overline{\mathrm{N}}(0_n, U_k)$ will yield a Gaussian predictive possibility function with expected value μ_k and variance $\tilde{\Sigma}_k + U_k$. Using Proposition 3, we can compute a map M_k from $\overline{\mathrm{N}}(\mu_k, \tilde{\Sigma}_k)$ to $\overline{\mathrm{N}}(\mu_k, \tilde{\Sigma}_k + U_k)$ and apply it to each particle $\tilde{x}_k^i$ to obtain the predicted ensemble at time k. The prediction step of the p-EnKF is summarised in Algorithm 4. Although we assume the Gaussianity of the ensemble at a slightly earlier stage than the standard EnKF, U_k is typically small compared to $\tilde{\Sigma}_k$ so that the impact of this assumption is expected to be small. In addition, fitting a Gaussian possibility function to the predicted ensemble is unnecessary as the result is known to be $\overline{\mathrm{N}}(\mu_k, \Sigma_k)$, with $\Sigma_k = \tilde{\Sigma}_k + U_k$, based on Proposition 1.

Algorithm 4 Prediction step of the p-EnKF at time k

Input: The time $k-1$ posterior ensemble, $\{(w^i, \hat{x}_{k-1}^i)\}_{i=0}^N$.
Output: The time k predictive ensemble, $\{(w^i, x_k^i)\}_{i=0}^N$, expected value μ_k, and variance Σ_k

  1: **for** $i \in \{1, \ldots, N+1\}$ **do**
  2:     $\tilde{x}_k^i \leftarrow F_k(\hat{x}_{k-1}^i)$ \\Apply dynamics to particles
  3: **end for**
  4: $\mu_k \leftarrow \tilde{x}_k^0$
  5: $\tilde{\Lambda}_k \leftarrow \Lambda^*\big(\{(w^i, \hat{x}_{k-1}^i)\}_i\big)$ \\Compute the precision matrix
  6: $\tilde{\Sigma}_k \leftarrow \tilde{\Lambda}_k^{-1}$
  7: **for** $i \in \{1, \ldots, N\}$ **do**
  8:     $T \leftarrow \big(\tilde{\Sigma}_k + U_k\big)^{1/2} \tilde{\Lambda}_k^{1/2}$ \\Compute matrix for adding uncertainty
  9:     $x_k^i \leftarrow \mu_k + T(\tilde{x}_k^i - \mu_k)$ \\Transport each particle
10: **end for**
11: $\Sigma_k \leftarrow \tilde{\Sigma}_k + U_k$

3.5 Update step

We first need to specify how the observation equation (1b) will be adapted to the considered context. Since perturbations in sensors are often stochastic in nature, we continue to model the error in the observation with a random variable, that is

$$Y_k = H_k(\boldsymbol{x}_k) + \varepsilon_k,$$

with $\varepsilon_k \sim \mathrm{N}(0, V_k)$ as before. The mechanism to update the information on $\boldsymbol{x}_k$ accordingly is provided by (2). In some situations, the main source of uncertainty in the observation will be of an epistemic nature, yet, if the corresponding model errors are described by $\overline{\mathrm{N}}(0, V_k)$, then the posterior possibility function will be the same; this follows from the likelihood principle since the two associated likelihoods will only differ by a multiplicative constant.

There are several variants of the probabilistic EnKF update step. Here, we follow the principles of the SqrtEnKF as it is well suited to the non-random setting of interest. In fact, we will show that our ensemble can be updated exactly in the same way as in the SqrtEnKF.

From the prediction step, we know that the best fitting Gaussian possibility function for the predicted sample is $\overline{\mathrm{N}}(\mu_k, \Sigma_k)$. As is standard, we consider the deviations $e_k^i = x_k^i - \mu_k$ and we verify that the correct updating formulas for the expected value and deviations are

$$\hat{\mu}_k = \mu_k + K_k(y_k - H_k\mu_k), \qquad \text{and} \quad \hat{e}_k^i = e_k^i - \tilde{K}_k H_k e_k^i,$$

where K_k and $\tilde{K}_k$ are as defined in the Kalman filter and SqrtEnKF, respectively. The updated particles obtained by adding the posterior expected value and deviations together are

$$\hat{x}_k^i = \hat{M}_k(x_k^i) \doteq (I_n - \tilde{K}_k H_k)x_k^i + (\tilde{K}_k H_k \mu_k + K_k y_k - K_k H_k \mu_k),$$

which is a linear transformation of x_k^i. Defining $\boldsymbol{x}_k$ as an uncertain variable described by $\overline{\mathrm{N}}(\mu_k, \Sigma_k)$ and using Proposition 3, it follows that $\hat{M}_k(\boldsymbol{x}_k)$ is described by

$$\begin{aligned}
\hat{f}_k &= \overline{\mathrm{N}}\left((I_n - \tilde{K}_k H_k)\mu_k + (\tilde{K}_k H_k \mu_k + K_k y_k - K_k H_k \mu_k), (I_n - \tilde{K}_k H_k)\Sigma_k(I_n - \tilde{K}_k H_k)^\top\right) \\
&= \overline{\mathrm{N}}\left(\mu_k + K_k(y_k - H_k\mu_k), (I_n - K_k H_k)\Sigma_k\right),
\end{aligned}$$

which matches the update mechanism of the KF, as required. The map $\hat{M}_k$ is therefore moving the particles in such a way that the best fitting Gaussian for $\{(w^i, \hat{M}_k(x_k^i))\}$ is the posterior of the KF with $\overline{\mathrm{N}}(\mu_k, \Sigma_k)$ as a predicted possibility function. Thus, the update step of the p-EnKF is formally the same as that of the standard SqrtEnKF, as detailed in Algorithm 2, albeit with a difference in interpretation.

3.6 Extensions

We finish this section by collecting together some extensions to the p-EnKF, demonstrating its applicability in the context of nonlinear measurement models and providing a number of ways to improve computational efficiency.

3.6.1 Nonlinear observation models

In the previous section, we have detailed the p-EnKF with a linear observation model. We now establish that, similarly to the EnKF [14, 33], the p-EnKF could be adapted to nonlinear observation models. There are, in fact, two main approaches to dealing with nonlinearity in the observation model, which we consider in turn.

Model linearisation The simplest way to handle a nonlinear observation model is to linearise it, i.e., to Taylor expand the observation model around the predictive expected value at time k as $H_k(x_k) \approx H_k(\mu_k) + J_{H_k}(\mu_k)(x_k - \mu_k)$, with $J_{H_k}(\mu_k)$ the Jacobian matrix of H_k at μ_k. Then, a new observation model can be defined based on the observation matrix $J_{H_k}(\mu_k)$ and on a non-zero mean $H_k(\mu_k) - J_{H_k}(\mu_k)\mu_k$ for the observation noise ε_k. After that, we can follow Algorithm 2 for the update, except that the term $y_k - H_k\mu_k$ is replaced by $y_k + J_{H_k}(\mu_k)\mu_k$ in step 4. Since linearisation does not depend on the chosen representation of uncertainty, it is equally applicable to the p-EnKF as it is to standard versions of the EnKF. The linearisation method can usually be improved by replacing the term $H_k \Sigma_k H_k^\top$ by the predictive variance of the observation based on the ensemble [33], which is also applicable to the p-EnKF. The term $\Sigma_k H_k^\top$ can usually also be replaced by a ensemble-based approximation, yet, this is not straightforward in the p-EnKF since the precision matrix does not have the same properties as the covariance matrix: to compute the covariance matrix $\Sigma_{\boldsymbol{x}, H_k(\boldsymbol{x})}$ between the uncertain variables $\boldsymbol{x}$ and $H_k(\boldsymbol{x})$, one must first compute the precision matrix for $(\boldsymbol{x}, H_k(\boldsymbol{x}))$, invert it, and then extract the block corresponding to $\Sigma_{\boldsymbol{x}, H_k(\boldsymbol{x})}$. Such an extension of the state is however usual, as described next.

Extending the state space Another way to deal with a nonlinear observation model is by extending/augmenting the original state with a the corresponding predicted observation, which is then observed linearly. In this case, the state becomes $\boldsymbol{z}_k = (\boldsymbol{x}_k, H_k(\boldsymbol{x}_k))$ and the extended observation matrix is $\tilde{H}_k = \begin{bmatrix} 0_{m \times n} & I_m \end{bmatrix}$, with $0_{m \times n}$ the 0 matrix of size $m \times n$. Algorithm 2 can be used as usual once a Gaussian is fitted to this extended state. The posterior ensemble can then be extracted by choosing the first n elements of the extended state for every particle. Care must be taken in practice as the precision matrix of the extended state can be close to singularity due to a strong correlation between the elements. For instance, if one of the components of the state that is observed independently becomes sufficiently well estimated at a given time step, then it might be that the nonlinear observation function is approximately linear from the viewpoint of the ensemble. Yet, this corresponds to cases where linearisation would be appropriate. It therefore appears that a hybrid technique would be the most suitable, with components of the observations being either linearised or included in the state depending on their observed degree of nonlinearity.

3.6.2 Techniques to improve computational efficiency

As with any ensemble-based technique, the accuracy of the p-EnKF depends on ensemble size. As it is typically of interest to use small ensembles for computational reasons, it can be challenging to represent the state of interest adequately. An important aspect is that the p-EnKF requires a minimum sample size equal to the state's dimension plus one due to the computation of the variance; a sample size matching the dimension plus one is sufficient to ensure that the resulting covariance matrix is of full rank providing only that the collection of displacements from the expected value to each of the sample points are linearly independent whereas a smaller sample will lead to a rank-deficient covariance matrix. That a sample size equal to the dimension plus one suffices follows from the fact that a quadratic form bounded away from zero at a number of points separated from the centroid by linearly-independent vectors cannot vanish anywhere.

However, the required sample size can be reduced by using one technique in the following section. Despite this constraint on the sample size, some methods can be considered for improving the computational efficiency, we present two of them in what follows.

Conditional Independence For the p-EnKF, variance is computed via an optimisation problem. Thus, the number of variables in the precision matrix of the state $x \in \mathbb{R}^n$ will be $n(n+1)/2$. However, for many problems, there is a natural structure in the state variable, such as a conditional-independence structure, which can be exploited to reduce this number. Indeed, for sparsely dependent models, it is straightforward to reduce the number of nonzero variables in the precision matrix Λ using conditional independence as follows: The off-diagonal elements Λ_{ij} that $i \neq j$ are set to 0 if the variable $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are to be modelled as conditionally independent. By setting some of the off-diagonal elements in the precision matrix to 0 during the computation of this matrix by optimisation, the other terms in the precision matrix will automatically adjust to these constraints, offering a systemic way to perform inflation that is tailored to the strength of the dependence being assumed away. However, exploiting this structure to gain in computational efficiency would require optimisation algorithms that are specifically tailored to sparse/band diagonal positive-definite matrices. The design of such algorithms is left for future work.

Nodal Numbering Scheme Apart from reducing the number of variables, the calculation and storage can be done more efficiently by reordering the variables to minimise the bandwidth of nonzero entries in the covariance matrix. One of the methods to achieve that is proposed in the literature is called the nodal numbering scheme [7]. This method uses the graph to reorder the state's element so that the nonzero elements will be close to the diagonal element. Moreover, the nodal numbering scheme ensures that the permuted matrix will have a bandwidth no greater than the original matrix, making the computation involved with the precision matrix more efficient, see [7] and [25] for more details.

4   Numerical Experiments

Here we show the performance of the p-EnKF using simulated data from two different models: a simple linear model and a modified Lorenz 96 model.

Data Generation One convenient feature of standard probabilistic modelling is that simulation can be performed exactly according to the model: the variability of scenarios, necessary for a thorough performance assessment, can be obtained directly by sampling from the assumed probability distributions. This is no longer the case with possibility theory since embracing epistemic uncertainty means that sampling is no longer a natural operation. The ideal solution would be to obtain a sufficiently large collection of real datasets for which the ground truth is known, this is however not generally achievable for problems like data assimilation. Instead, we generate our simulated scenarios by sampling from the probability distributions assumed by the probabilistic baselines and align our possibilistic model with these.

4.1   Linear model

We first consider a linear model since the performance can be clearly compared with the optimal filter, which can be computed in this instance via the KF. A simple linear model is considered so as to generalise easily to arbitrary dimensions. In particular, we consider that the i-th component at time k, $i \in \{2, \ldots, n\}$, only depends on the i-th and $(i-1)$-th components at time $k-1$. This means that conditional independence can be imposed to reduce the computational cost of obtaining the precision matrix with a limited information loss. The model can be written as a state-space model (1), for $k \in \{1, \ldots, 100\}$, with the following components:

1. The initial state X_0 is sampled from $\mathrm{N}(0_n, 10I_n)$.

2. The dynamic model is linear: $F_k(X_{k-1}) = F_k X_{k-1}$ with

$$F_k = \begin{bmatrix} 1 & \lambda & 0 & \cdots & 0 \\ 0 & 1 & \lambda & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix},$$

where $\lambda = 0.1$ and the covariance matrix of the dynamical noise ϵ_k is $U_k = 0.01 I_n$.

3. The observation model is also linear, $H_k(X_k) = H_k X_k$, with $H_k = \begin{bmatrix} I_m & 0_{m \times (n-m)} \end{bmatrix}$ and the covariance matrix of the observation noise ε_k is $V_k = 0.1 I_m$.

Based on all the shared properties between the Gaussian possibility function and the Gaussian distribution, the best way to align our possibilistic model with the assumed probabilistic one is to simply keep the expected value and variance parameters in our possibility functions. In particular, we assume that the initial state $\boldsymbol{x}_0$ is described by the Gaussian possibility function $\overline{N}(0_n, 10 I_n)$ and that the errors in the dynamical model are described by $\overline{N}(0_n, U_k)$.

The parameters of the considered methods are as follows: Unless otherwise stated, the parameter N is set to twice the state's dimension, i.e., $N = 2n$. For a given value of N, the actual number of samples for all methods is $N + 1$. The parameters of the UKF chosen in this paper are given by $\alpha = 0.25, \kappa = 130, \lambda = \alpha^2(n + \kappa) - n$ and $\beta = 2$.
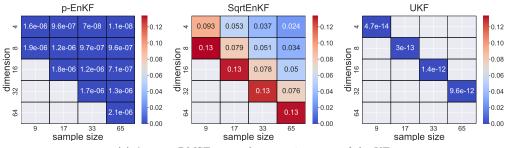
4.1.1 Performance assessment for various sample sizes and dimensions

Since the p-EnKF is an ensemble-based method, it is natural for us to investigate the performance based on different sample sizes and dimensions first. Figure 2 shows the performance of the p-EnKF with no localisation when the state is fully observed ($n = m$), comparing it to the SqrtEnKF and the UKF. The performance is measured in terms of average RMSE over 1000 realisations, except for $n = 64$ where we only consider 50 realisations due to a large computational time (more than 30 minutes per run). We aim to assess the performance after initialisation and thus focus on the estimation of the state X_{100} at the last time step. The RMSE is computed i) for the posterior expected value with respect to (w.r.t.) the posterior mean of the KF, ii) for the posterior expected value w.r.t. the true state, and iii) for the posterior variance w.r.t. the posterior variance of the KF. The key aspects in Figure 2 are as follows:
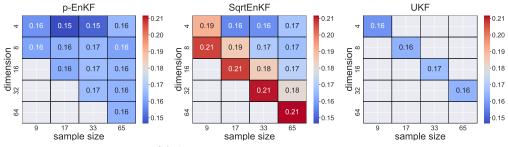
1. As can be seen in Figure 2a, despite the similarities between the p-EnKF and the SqrtEnKF, the latter improves on the former by at least 4 orders of magnitude in terms of RMSE w.r.t. the posterior mean of the KF. Although the SqrtEnKF could be used when $N < n$, which is key in large dimensions, the performance would necessary be lower in this regime. The difference in performance is still visible when considering the RMSE w.r.t. the true state, as in Figure 2b, however it is less pronounced due to the unavoidable error caused by the distance between the true state and the optimal estimator given by the mean of the KF.

2. Despite the fact that the capabilities in terms of variance recovery are almost indistinguishable in Figure 1a between the possibilistic and probabilistic approach, Figure 2c shows that the p-EnKF once again largely outperforms the SqrtEnKF in terms of RMSE w.r.t. the optimal variance given by the KF, with improvements by at least 4 orders of magnitude throughout once again. This is due to the fact that here, as opposed to Figure 1a, the mean of the ensemble also needs to be estimated by the SqrtEnKF whereas the expected value for the p-EnkF is given by the particle with index 0 and thus does not need to be re-estimated.

The estimates of the UKF are closer to optimality than the ones of the p-EnKF in Figures 2a and 2c; this could be due to the difference in initialisation between the two algorithms, with the UKF placing points deterministically and with the p-EnKF relying on a random sample at the first time step. Yet, this source
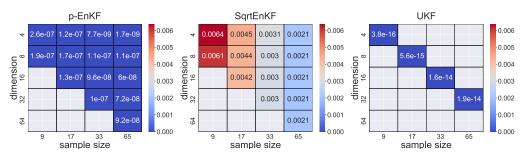
of randomness in the p-EnKF is not necessary and other initialisation schemes are considered in the next section.

(a) Average RMSE w.r.t. the posterior mean of the KF.

(b) Average RMSE w.r.t. the true state.

(c) Average RMSE w.r.t. the posterior variance of the KF.

Figure 2: Performance assessment for the fully-observed linear model.

4.1.2 Alternatives schemes for particle initialisation

We now investigate the performance of the p-EnKF with different initialisation schemes inspired by the UKF. When $N = 2n$, we can simply consider exactly the σ-points of the UKF as initial point for the p-EnKF and we refer to this scheme as "UKF initialisation". To allow for setting $N = n$, we also arbitrarily consider only the σ-points which corresponds to increasing (resp. decreasing) one of the components of the mean vector and we refer to this scheme as "UKF initialisation $+$" (resp. "UKF initialisation $-$").

To better highlight the dependency on the initialisation scheme, we consider in this section a partially observed model with $m = 1$, i.e., only the first component of the state is observed. This is challenging in general, so only problems of small state dimension are considered. In Figure 3, the performance assessment is carried out for $n = 3$ and $n = 5$, with all RMSEs being averaged over 1000 repeats. In Figures 3a

and 3b, all the considered initialisation schemes are compared against the SqrtEnKF and the StEnKF with $2n+1$ samples, with the poor performance of the latter highlighting the difficulty of these inference problems despite the small dimension. There is a slight but consistent improvement in performance when using the UKF initialisation schemes in the p-EnKF, although this improvement vanishes after 30 to 50 time steps, depending on the state dimension. Figures 3c and 3d are restricted to the StEnkKF, StEnKF and p-EnKF with UKF initialisation for the sake of legibility, and show that although the interquartile range of the different methods overlap, the p-EnKF performs consistently well across different realisations.

(a) Average RMSE w.r.t. the true state for $n = 3$

(b) Average RMSE w.r.t. the true state for $n = 5$

(c) Average RMSE with the interquartile range for $n = 3$

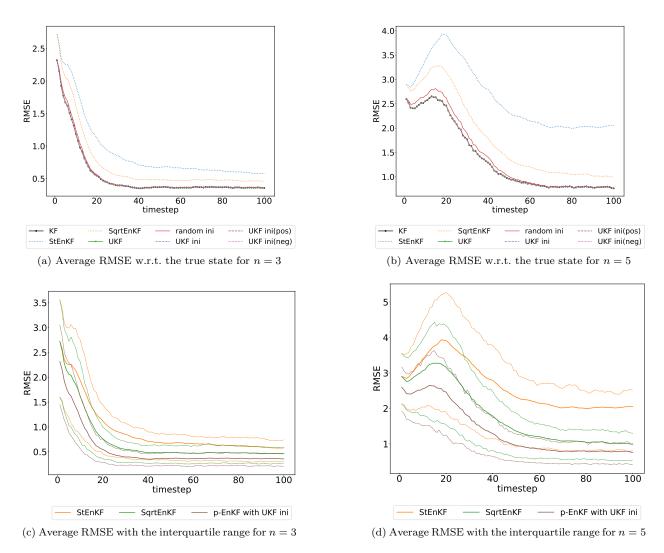(d) Average RMSE with the interquartile range for $n = 5$

Figure 3: Average RMSE and the error range of different algorithms in the linear model. Left: state's dimension of 3; Right: state's dimension of 5, averaged over 1000 repeats.

4.1.3 Banded vs. full precision matrix

To illustrate the capabilities of the p-EnKF to deal with localisation via a systematic form of inflation, we contrast its performance with a full precision matrix against the one with a banded precision matrix with a bandwidth of two, i.e., where all elements except the diagonal and elements adjacent to it are set to zero. We compare these results against the ones obtained with the standard versions of the EnKF for

the fully-observed ($n = m = 5$) and partially-observed ($n = 5, m = 1$) cases. The performance is measured with three quantities i) the RMSE of the posterior expected value w.r.t. the true state, ii) the determinant of the posterior variance, and iii) the Mahalanobis distance between the posterior expected value and the true state. The Mahalanobis distance is used to capture how good the estimate is relative to its variance and, hence, assesses the calibration of the algorithms in terms of uncertainty quantification. It is defined as $\sqrt{(x - \mu)^\top \Sigma^{-1}(x - \mu)}$ where x is the true state, μ is a given posterior expected value, and Σ is the posterior variance.

Figure 4 shows the performance of the p-EnKF with full and banded precision matrix against the same baselines as before. All the algorithms except the KF use 11 samples, i.e., the ensemble size is $N + 1$ where $N = 2n$ and the considered performance metrics are averaged over 1000 repeats. The important aspects in Figure 4 are as follows:

1. As can be seen in Figures 4c and 4d, forcing the precision matrix to be banded has little impact on the precision of the p-EnKF, despite correlations being crucial for a strong performance in the partially observed case (Figure 4d). This is confirmed in Figures 4c and 4d, where the log-determinant of the posterior variance is mostly unaffected by localisation. A small but noticeable difference can be seen in Figure 4d, but the change in variance is in the correct direction: the determinant of the variance was increased by localisation, i.e., some inflation has been automatically applied in order to compensate for the imposed conditional independence.

2. The Mahalanobis distance for the two versions of the the p-EnKF is nearly constant and close to the one of the KF for both considered scenario, as seen in Figures 4e and 4f. Conversely, the SqrtEnKF and StEnKF both display large Mahalanobis distances, with the one of the StEnKF even diverging in the partially-observed case; this is due to these algorithms having a larger RMSE than the KF (Figures 4c and 4d) but a smaller variance (Figures 4c and 4d).

In conclusion, the p-EnKF adopts a nearly optimal behaviour in this linear scenario despite partial observability and localisation. In contrast, the standard versions of the EnKF depart significantly from the behaviour of the KF and tend to be overly optimistic even in the considered small-dimensional inference problems, with well known adverse consequences for downstream tasks for which a reliable quantification of the uncertainty can be crucial.

4.2 Modified Lorenz 96 type model

In order to show that the strong performance of the p-EnKF observed in the previous section generalises beyond the linear case, we now consider a modified Lorenz 96 (LR96) model, which can be written as a state-space model (1), with $k \in \{1, \ldots, 100\}$, with the following components:

1. The initial state X_0 is sampled from $N(0_n, 10 I_n)$.

2. The deterministic part of the dynamic model is characterised by $x = F_k(x')$ with

$$
\begin{aligned}
x_1 &= x_1' + \left((x_2' - c)c - x_1' + F\right)\Delta t \\
x_2 &= x_2' + \left((x_3' - c)x_1' - x_2' + F\right)\Delta t \\
x_i &= x_i' + \left((x_{i+1}' - x_{i-2}')x_{i-1}' - x_i' + F\right)\Delta t \qquad \text{for } 3 \leq i \leq n - 1 \\
x_n &= x_n' + \left((c - x_{n-2}')x_{n-1}' - x_n' + F\right)\Delta t
\end{aligned}
$$

where "x_i" refers to the i-th component of x and where $F = 8$, $c = 1$, and $\Delta t = 0.01$.

3. The covariance matrix of the dynamical noise ϵ_k is $U_k = 0.01 I_n$.

4. The observation model is linear, $H_k(X_k) = H_k X_k$, with $H_k = \begin{bmatrix} I_m & 0_{m \times (n-m)} \end{bmatrix}$ and the covariance matrix of the observation noise ε_k is $V_k = 0.1 I_m$.
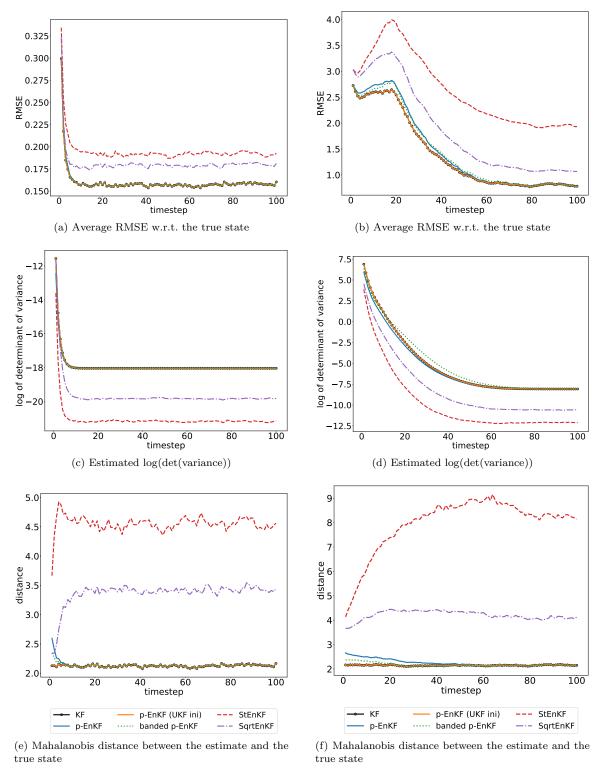
(a) Average RMSE w.r.t. the true state

(b) Average RMSE w.r.t. the true state

(c) Estimated log(det(variance))

(d) Estimated log(det(variance))

(e) Mahalanobis distance between the estimate and the true state

(f) Mahalanobis distance between the estimate and the true state

Figure 4: Performance assessment for the linear model with $n = 5$ when (left) fully observed, $m = 5$ and (right) partially observed, $m = 1$, averaged over 1000 repeats.

Overall, the only difference with the linear model is the transition function F_k. Throughout this section, the ensemble size will be set to $2n + 1$, that is $N = 2n$, unless specified otherwise. The LR96 model is particularly convenient for performance assessment since the dimension can be easily adjusted.

As before, we start by investigating the performance of the p-EnKF with the full precision matrix based on different sample sizes and dimensions such that all elements of the state are observed ($n = m$) compared to the SqrtEnKF and the UKF. However, for the nonlinear model, we only examine the performance in terms of RMSE w.r.t. the true state at the last step since we can no longer obtain the optimal state estimate from the KF. The RMSE displayed in Figure 5 is averaged over 100 repeats, except for $n = 64$ where we only consider 50 realisations due to the computational cost. The results shown in Figure 5 are very close to the ones obtained in the linear case, with all three algorithms maintaining a similar level of performance despite the non-linearities.
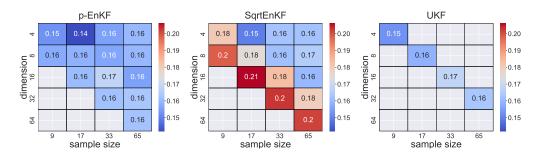

Figure 5: Performance of p-EnKF in terms of RMSE w.r.t. the true state, compared with the SqrtEnKF and the UKF for the LR96 model

Because of the similarities between the results with the linear model and the LR96 model, we only highlight where noticeable differences arise. Figure 6a shows that, in the partially-observed case with $m = 1$ and $n = 5$, the SqrtEnKF has a RMSE that is now closer to the StEnKF than to the p-EnKF. Localisation in the p-EnKF still has a very mild effect, although the error increases around time step 70. This increase in error is however captured by the associated covariance matrix, so that the Mahalanobis distance remains constant throughout the scenario, as required. This behaviour suggests that the correlation increased around time step 70, forcing the p-EnKF with banded matrix to increase the amount of inflation to compensate for the loss of information caused by the imposed conditional independence.
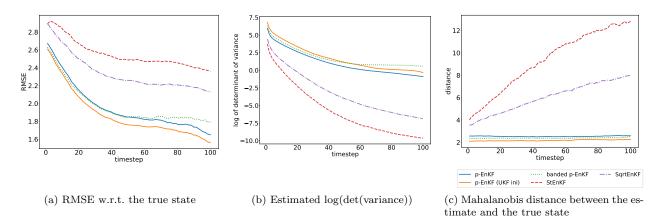

(a) RMSE w.r.t. the true state (b) Estimated log(det(variance)) (c) Mahalanobis distance between the estimate and the true state

Figure 6: Performance for a partially-observed LR96 model ($m = 1$) when $n = 5$, averaged over 1000 repeats.

20

5 Conclusion

We have introduced the possibilistic ensemble Kalman filter, or p-EnKF, a data assimilation technique treating the state of a state-space model as a fixed quantity about which limited information is available. By using possibility theory to model this form of epistemic uncertainty, we found that much of the intuition behind the standard versions of the EnKF remains valid, with the differences between the theories of possibility and probability leading to key features in the p-EnKF. Specifically, the properties of the expected value and variance in possibility theory appeared to be beneficial for inference problems of small to moderate dimensions, with the p-EnKF closely approximating the Kalman filter in the linear-Gaussian case. These properties also allowed for localisation to be seamlessly applied with no parameter tuning required to compensate for the loss of information incurred by the imposed conditional independence.

In the current version of the p-EnKF, the computation of covariance matrices relies on solving a constrained optimisation problem, which is time-consuming and requires an ensemble size greater than the dimension of the state. Although beyond the scope of this work, lifting these constraints appears to be feasible through the use of specialised optimisation techniques and suitable regularisation.

There are also several possible avenues for further investigation. We highlight two directions that we think are immediately interesting: although we have found the algorithms performance to be robust to the specification of the initial ensemble, there is a potential for developing systematic approaches to specifying it which might lead to further improved performance with small ensembles; and, an open question is how can the update step be reformulated to operate directly in terms of the precision matrix to avoid the need for matrix inversion, which would be required to facilitate the use of the p-EnKF in high-dimension.

Acknowledgements

References

[1] Rossella Arcucci, Lamya Moutiq, and Yi-Ke Guo. Neural assimilation. In *Computational Science–ICCS 2020: 20th International Conference, Amsterdam, The Netherlands, June 3–5, 2020, Proceedings, Part VI 20*, pages 155–168. Springer, 2020.

[2] Mark Asch, Marc Bocquet, and Maëlle Nodet. *Data assimilation: methods, algorithms, and applications*. SIAM, 2016.

[3] Cédric Baudrit, Didier Dubois, and Nathalie Perrot. Representing parametric probabilistic models tainted with imprecision. *Fuzzy sets and systems*, 159(15):1913–1928, 2008.

[4] Adrian N Bishop and Pierre Del Moral. Robust Kalman and Bayesian set-valued filtering and model validation for linear stochastic systems. *SIAM/ASA Journal on Uncertainty Quantification*, 11(2):389–425, 2023.

[5] Gerrit Burgers, Peter Jan Van Leeuwen, and Geir Evensen. Analysis scheme in the ensemble Kalman filter. *Monthly weather review*, 126(6):1719–1724, 1998.

[6] Zhijin Chen, Branko Ristic, Jeremie Houssineau, and Du Yong Kim. Observer control for bearings-only tracking using possibility functions. *Automatica*, 133:109888, 2021.

[7] Elizabeth Cuthill and James McKee. Reducing the bandwidth of sparse symmetric matrices. In *Proceedings of the 1969 24th National Conference*, pages 157–172, 1969.

[8] Bernard De Baets, Elena Tsiporkova, and Radko Mesiar. Conditioning in possibility theory with strict order norms. *Fuzzy Sets and Systems*, 106(2):221–229, 1999.

[9] Arthur P Dempster. The Dempster–Shafer calculus for statisticians. *International Journal of Approximate Reasoning*, 48(2):365–377, 2008.

[10] Didier Dubois, Hung T Nguyen, and Henri Prade. Possibility theory, probability and fuzzy sets misunderstandings, bridges and gaps: Misunderstandings, bridges and gaps. In *Fundamentals of fuzzy sets*, pages 343–438. Springer, 2000.

[11] Didier Dubois and Henry Prade. Possibility theory and its applications: Where do we stand? *Springer Handbook of Computational Intelligence*, pages 31–60, 2015.

[12] Paul Fearnhead and Hans R Künsch. Particle filters and data assimilation. *Annual Review of Statistics and Its Application*, 5:421–449, 2018.

[13] Ronald A Fisher. The fiducial argument in statistical inference. *Annals of Eugenics*, 6(4):391–398, 1935.

[14] Marco Frei. *Ensemble Kalman filtering and generalizations*. PhD thesis, ETH Zurich, 2013.

[15] Philipp A Guth, Vesa Kaarnioja, Frances Y Kuo, Claudia Schillings, and Ian H Sloan. A quasi-Monte Carlo method for optimal control under uncertainty. *SIAM/ASA Journal on Uncertainty Quantification*, 9(2):354–383, 2021.

[16] Jan Hannig, Hari Iyer, Randy CS Lai, and Thomas CM Lee. Generalized fiducial inference: A review and new results. *Journal of the American Statistical Association*, 111(515):1346–1361, 2016.

[17] Jeremie Houssineau. Parameter estimation with a class of outer probability measures. *arXiv preprint arXiv:1801.00569*, 2018.

[18] Jeremie Houssineau. A linear algorithm for multi-target tracking in the context of possibility theory. *IEEE Transactions on Signal Processing*, 69:2740–2751, 2021.

[19] Jeremie Houssineau and Adrian N Bishop. Smoothing and filtering with a class of outer measures. *SIAM/ASA Journal on Uncertainty Quantification*, 6(2):845–866, 2018.

[20] Jeremie Houssineau, Neil K Chada, and Emmanuel Delande. Elements of asymptotic theory with outer probability measures. *arXiv preprint arXiv:1908.04331*, 2019.

[21] Tijana Janjić, Lars Nerger, Alberta Albertella, Jens Schröter, and Sergey Skachko. On domain localization in ensemble-based Kalman filter algorithms. *Monthly Weather Review*, 139(7):2046–2060, 2011.

[22] R. Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82:35–42, 1960.

[23] Eugenia Kalnay. *Atmospheric modeling, data assimilation and predictability*. Cambridge University Press, 2003.

[24] Matthias Katzfuss, Jonathan R Stroud, and Christopher K Wikle. Understanding the ensemble Kalman filter. *The American Statistician*, 70(4):350–357, 2016.

[25] Chatchuea Kimchaiwong. *Ensemble Kalman filtering with an alternative representation of uncertainty*. PhD thesis, University of Warwick, 2024.

[26] Christiane Lemieux. *Monte Carlo and quasi-Monte Carlo sampling*, volume 20. Springer, 2009.

[27] Judea Pearl. Reasoning with belief functions: An analysis of compatibility. *International Journal of Approximate Reasoning*, 4(5-6):363–389, 1990.

[28] Ruth E Petrie. Localization in the ensemble Kalman filter. *MSc Atmosphere, Ocean and Climate University of Reading*, 2008.

[29] Sebastian Reich and Colin Cotter. *Probabilistic forecasting and Bayesian data assimilation*. Cambridge University Press, 2015.

[30] Christian P Robert and George Casella. *Monte Carlo statistical methods*, volume 2. Springer, 1999.

[31] Simo Särkkä and Lennart Svensson. *Bayesian filtering and smoothing*. Cambridge University Press, 2023.

[32] Amirhossein Taghvaei and Prashant G Mehta. An optimal transport formulation of the ensemble Kalman filter. *IEEE Transactions on Automatic Control*, 66(7):3052–3067, 2020.

[33] Youmin Tang, Jaison Ambandan, and Dake Chen. Nonlinear measurement function in the ensemble Kalman filter. *Advances in Atmospheric Sciences*, 31(3):551–558, 2014.

[34] Gabriel A Terejanu et al. Extended Kalman filter tutorial. *University at Buffalo*, 2008.

[35] Michael K Tippett, Jeffrey L Anderson, Craig H Bishop, Thomas M Hamill, and Jeffrey S Whitaker. Ensemble square root filters. *Monthly weather review*, 131(7):1485–1490, 2003.

[36] Peter Jan Van Leeuwen, Yuan Cheng, and Sebastian Reich. *Nonlinear data assimilation for high-dimensional systems: -with geophysical applications*. Springer, 2015.

[37] Eric A Wan and Rudolph Van Der Merwe. The unscented Kalman filter for nonlinear estimation. In *Proceedings of the IEEE 2000 Adaptive Systems for Signal Processing, Communications, and Control Symposium (Cat. No. 00EX373)*, pages 153–158. IEEE, 2000.

[38] Jeffrey S Whitaker and Thomas M Hamill. Ensemble data assimilation without perturbed observations. *Monthly Weather Review*, 130(7):1913–1924, 2002.