

On the guarantees derived from a possibilistic interpretation of ensemble predictions and their operational use

Noémie Le Carrer

Institute for Risk and Uncertainty, Dep. of Engineering, University of Liverpool, United Kingdom

Abstract.

Ensemble forecasting has become popular in weather prediction to reflect the uncertainty about high-dimensional, nonlinear systems with extreme sensitivity to initial conditions. By means of small strategical perturbations of the initial conditions, sometimes accompanied with stochastic parameterisation schemes of the atmosphere-ocean dynamical equations, ensemble forecasting aims at sampling possible future scenarii and ideally at interpreting them in a Monte-Carlo-like approximation. Traditional probabilistic interpretations of ensemble forecasts do not take epistemic uncertainty into account, nor the fact that ensemble predictions cannot always be interpreted in a density-based manner due to the strongly nonlinear dynamics of the atmospheric system. As a result, probabilistic predictions are not always reliable, especially in the case of extreme events. In this work, we investigate whether relying on possibility theory can circumvent these limitations. We show how it can be used to compute confidence intervals with guaranteed reliability, when a classical probabilistic postprocessing technique fails to do so in the case of extreme events. We empirically illustrate our approach with an imperfect version of the Lorenz 96 model, and demonstrate that it is promising for risk-averse decision-making.

Keywords: Possibility theory; Decision-making under uncertainty; Reliability; Statistical post-processing; Extreme event prediction; Imprecise probabilities; Risk-averse decision-making

1. Introduction

In weather forecasting, it is acknowledged that by design (limited size of set of ensemble predictions — EPS, targeted sampling of initial conditions — ICs) and by context (flow-dependent regime error, strongly nonlinear system), raw ensemble forecasts generally do not provide reliable probabilistic predictions (Bröcker and Smith, 2008; Gneiting and Katzfuss, 2014). This is especially the case for extreme events (Legg and Mylne, 2004). The latter result from nonlinear interactions at small scales, which implies that they generally cannot be associated with a high density of ensemble members (Mylne et al., 2002). Ensemble forecasts are made more reliable and operational via recalibration (Buizza, 2018), whose aim can be summarized as "finding the transformation that, applied to the raw ensemble, leads to the probability distribution that will maximise a performance metric on a training set". In spite of the diversity of approaches developed in the literature (Buizza, 2018) and their technical success for improving the prediction skills when it comes to common events, the actionability of probabilistic predictions often remains problematic (Smith, 2016). In

particular, the probabilistic prediction of extreme events often needs a development on its own (Friederichs and Hense, 2007; Friederichs et al., 2018).

(Bröcker and Smith, 2008) questioned whether probability distributions constitute *the best representation of the valuable information contained in an EPS*. We advance convincing arguments that possibility theory, “a weaker theory than probability [...] also relevant in non-probabilistic settings where additivity no longer makes sense” (Dubois et al., 2004), is an interesting alternative. Our investigation is particularly relevant since conceptual and practical limitations restrict the applicability of a density-based (i.e. additive) interpretation of EPSs. We show how interpreting EPSs in a possibilistic way brings useful formal guarantees on the derived confidence intervals, even in the case of extreme events.

Section 2 summarises the basics of possibility theory, Section 3 presents our possibilistic framework and discusses the theoretical guarantees that can be associated with its outputs. Section 4 introduces the synthetic experiments on the Lorenz 96 system (Lorenz, 1996) (L96) which allow us to assess these guarantees and their operational cost for both common and extreme events. We compare them with the outputs of a classical probabilistic interpretation of EPSs, and discuss our results in Section 5.

2. Possibility theory

Possibility theory is an uncertainty theory developed from fuzzy set theory by (Zadeh, 1978) and (Dubois and Prade, 2012). It is designed to handle incomplete information and represent ignorance. Considering a system whose state is described by a variable $x \in \mathcal{X}$, the possibility distribution $\pi : \mathcal{X} \rightarrow [0, 1]$ represents the available information about the current state of the system. Given an event $A = \{x \in S_A\}$, where S_A is a subset of $\mathcal{X}$, the possibility and necessity measures are defined respectively as: $\Pi(A) = \sup_{x \in S_A} \pi(x)$ and $N(A) = 1 - \Pi(\bar{A})$ where $\bar{A}$ represents the complementary

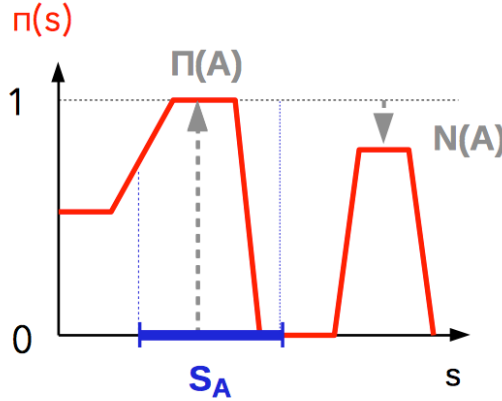


Figure 1. Possibility distribution $\pi(s)$ where for an event of interest $A = "s \in S_A"$, the possibility $\Pi(A)$ and necessity $N(A) = 1 - \Pi(\bar{A})$ measures are represented.

A possibilistic interpretation of ensemble predictions

event of A (see Figure 1 for a visual understanding of these quantities). Both measures satisfy the following axioms and conventions (Cayrac et al., 1994):

1. $\Pi(\mathcal{X}) = 1$ and $\Pi(\emptyset) = 0$
2. $\Pi(A \cup B) = \max(\Pi(A), \Pi(B))$
3. $N(A) = 1 \Leftrightarrow \Pi(\bar{A}) = 0$ indicates that A has to happen, it is necessary: $\bar{A}$ is impossible;
4. $0 < N(A) < 1$ is a tentative acceptance of A to a degree $N(A)$;
5. $(\Pi(A) = \Pi(\bar{A}) = 1) \Leftrightarrow (N(A) = N(\bar{A}) = 0)$ represents total ignorance: the evidence doesn't allow us to conclude whether A is rather true or false.

Possibility and probability distributions are interconnected through the concept of imprecise probabilities (Dempster, 2008). A probability measure P and possibility measure Π are consistent iff (Dubois et al., 2004):

$$P(A) \leq \Pi(A), \quad \forall A \quad (1)$$

The definition of necessity implies that in these conditions:

$$N(A) \leq P(A) \leq \Pi(A), \quad \forall A \quad (2)$$

.

2.0.0.1. From data to possibility distribution Let $x \in \mathcal{X}$ be a stochastic variable for which we try to make a prediction. The evidence about x_t is a set $S = \{x_1, \dots, x_{N_s}\}$ of N_s samples of x , which we assume has been randomly generated from an unknown probability distribution P . To turn this information into a possibility distribution describing the knowledge on the actual value of x , we use the technique developed by (Masson and Denceux, 2006). Their methodology is specifically designed to derive a possibility distribution from scarce data. The idea is, after binning the x -axis into n bins, to recover the simultaneous confidence intervals at level β on the true probability $P(x \in b_i)$ for each bin b_i . From these confidence intervals and considerations about Equation (1), the procedure allows us to compute a possibility distribution $\pi(x)$ that *dominates* with confidence β the true probability distribution (i.e. Equation (1) is verified in $100\beta\%$ of the cases). The simultaneous confidence intervals for multinomial proportions are computed by means of Goodman's formulation (Goodman, 1965). This procedure takes into account the uncertainty on the multinomial proportions that is due to the limited size of S . This is fundamental for our application, which is to seek guarantees on the possibility of observing a given event.

As shown by Equation (2), a possibility distribution can be seen as a complete and consistent framework to deal with imprecise probabilities. Although the above procedure for computing a possibility distribution mostly relies on probabilities, its result contains more information than a purely probabilistic distribution *in the situation of incompleteness* (typically implied by a small dataset S). Indeed, the interval on the true probability allows the incompleteness of data or knowledge to be accounted for, while a point probability hides the fact that the said probability cannot be fully trusted (e.g. due to epistemic uncertainty). Figure 2 illustrates the results of this

Noémie Le Carrer

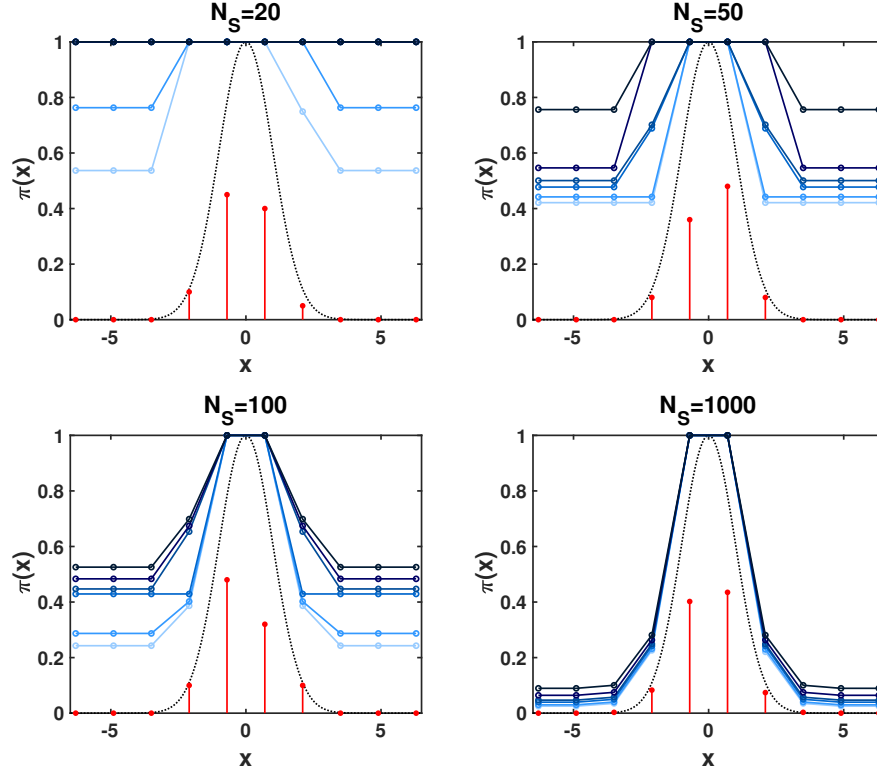


Figure 2. Possibility distributions (solid lines) derived from datasets of N_s elements sampled from a standard normal distribution. This derivation requires the computation of simultaneous confidence intervals for multinomial proportions over the x -axis binned into $n = 10$ bins. The effect of the confidence level $\beta = \{0.6, 0.75, 0.9, 0.95, 0.99, 1\}$ of the Goodman's formulation is shown (larger β are plotted darker). Vertical red lines represent a frequency histogram of the same datasets and the normalised underlying Gaussian distribution is represented as a dotted line.

methodology applied to datasets sampled from a normal distribution, for various levels of β and N_s . For a given N_s , the larger β is, the more conservative is the distribution: γ such as $\pi(x) \geq \gamma \forall x$ is larger, which implies that for any event $A \subset \mathcal{X}$: $\Pi(\bar{A}) \geq \gamma$. This also reads: $N(A) \leq 1 - \gamma$, meaning that the confidence level associated with any A cannot reach high values. Increasing N_s reduces the relative effect of β and all distributions tend in shape towards the underlying probability distribution, even if the tails remains more conservative for larger β .

3. Proposed framework

We are interested in the prediction of the state variable x_{t_0+t} of a dynamical system at lead time t , starting from the IC x_{t_0} . For simplicity, we omit the reference to t_0 and note x_t the *verification*. In the EPS context, given a numerical prediction model $\mathcal{M}$, the elements of information at hand are:

1. An ensemble of M predictions at lead time t , the ensemble members or EPS, obtained by means of $\mathcal{M}$ applied to slightly perturbed ICs around t_0 : $\tilde{\mathbf{x}}_t = \{\tilde{x}_t^1, \dots, \tilde{x}_t^M\}$.

A possibilistic interpretation of ensemble predictions

2. An archive $\mathcal{I}_t$ containing the pairs $(\tilde{\mathbf{x}}_{t_0+t}, x_{t_0+t})$ for the lead time t of interest and N_I different instances of t_0 . These instances are chosen so that the initial points of two successive trajectories are statistically independent from each other.

3.1. DERIVING POSSIBILITY DISTRIBUTIONS FROM EPSs

The objective of our possibilistic interpretation of EPSs is to derive from an EPS $\tilde{\mathbf{x}}_t$ and the archive $\mathcal{I}_t$ a possibility distribution $\pi(x_t|\tilde{\mathbf{x}}_t, \mathcal{I}_t)$, that encodes the knowledge derived from $\tilde{\mathbf{x}}_t$ about the verification x_t . The procedure described in this section is summarised and illustrated in the steps 1—5 of Figure 3.

Both system and model being (to a certain extent) deterministic and (close to) stationary, the past behaviour of the couple {system, model} is representative of its future behaviour. Consequently, if we are able to enumerate the possible values (already seen in $\mathcal{I}_t$ or not) for the verification x_t associated with a small range S_{x_t} of the values taken by ensemble members, then a future observation x_t should belong to that set of possible values when an ensemble member $\tilde{x}_t^m$ falls within S_{x_t} . Beyond that, we would like to know which ones of these values are more possible than others for x_t . In other words, we want to estimate the possibility distribution $\pi(x_t|\tilde{x}_t^m \in S_{x_t})$. Because there is no notion of ‘density’ of the evidence in the possibilistic perspective (at least in our rationale for choosing this framework), the number of ensemble members falling in S_{x_t} will not affect the resulting possibility distribution for x_t .

To make use of the full set of ensemble members, we first partition the x -axis into n bins b_i , take the subset B of bins occupied by at least one ensemble member of the EPS, and compute $|B|$ possibility distributions $\pi(x_t|\tilde{x}_t^m \in b_j)$ where $b_j \in B$. Namely, for each bin $b_j \in B$ occupied by at least one ensemble member $\tilde{x}_t^m \in \tilde{\mathbf{x}}_t$, we retrieve the N_s ensemble members $\tilde{x}_t^m \in b_j$ in the archive $\mathcal{I}_t$ and build a histogram of the set of corresponding verifications (so-called *analogs*) over the same binned x -axis. We then derive $\pi(x_t|\tilde{x}_t^m \in b_j)$ following the methodology presented in Section 2.0.0.1.

We obtain $|B|$ possibility distributions $\pi(x_t|\tilde{x}_t^m \in b_j)$, each dominating with confidence β the true probability distribution $P(x_t|\tilde{x}_t^m \in b_j)$. Each possibility distribution provides the possibilities for the verification x_t given the presence of one or more ensemble members in bin b_j and is thus a partial view on the state x_t . Since there is only one truth for x_t , we can merge them through a union operator. Fuzzy set theory offers several definitions for computing the distribution resulting from the union of two fuzzy distributions. We adopt here the standard definition for its intuitive rationale: $\pi_{A \cup C}(x) = \max(\pi_A(x), \pi_C(x))$.

We construct the resulting possibility distribution as:

$$\pi_{EPS}(x_t \in b_i|\tilde{\mathbf{x}}_t) = \cup_{j|b_j \in B} \pi(x_t|\tilde{x}_t^m \in b_j) = \sup_{j|b_j \in B} \pi(x_t \in b_i|\tilde{x}_t^m \in b_j), \quad i = 1, \dots, n. \quad (3)$$

3.2. FROM POSSIBILITY DISTRIBUTION TO PREDICTION

We focus on the continuous interpretation of π_{EPS} and now turn to our approach for producing confidence intervals for the future value x_t , and on the associated formal guarantees.

Noémie Le Carrer

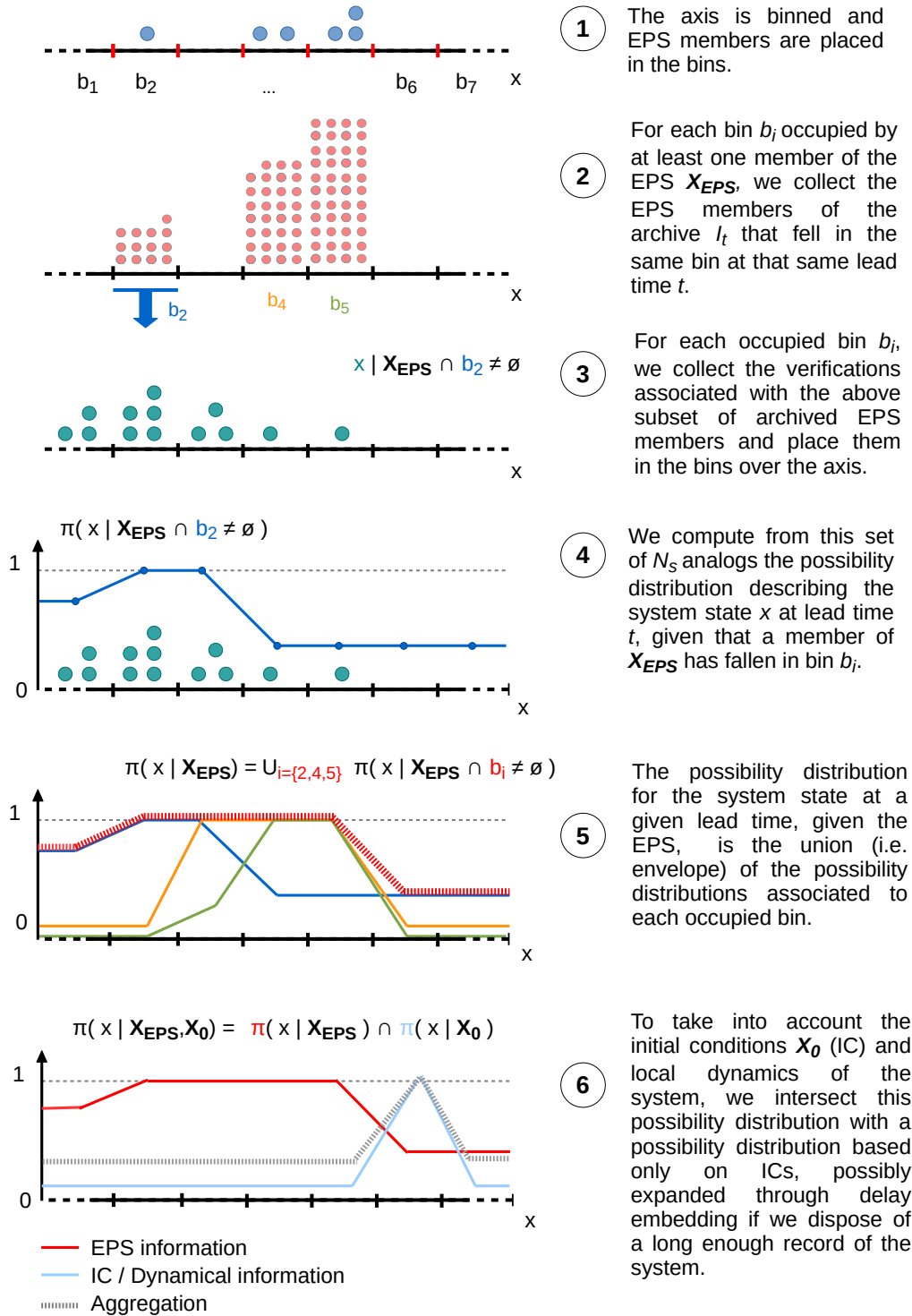


Figure 3. Methodology of the possibilistic interpretation of EPSs developed in this paper.

A possibilistic interpretation of ensemble predictions

As can be easily derived from Equation (2), a possibility density π is consistent with the associated probability measure P iff its α -cuts $C_\pi^\alpha = \{x, \pi(x) \geq \alpha\}$ satisfy:

$$P(x \in C_\pi^\alpha) = P(C_\pi^\alpha) \geq 1 - \alpha, \forall \alpha \in [0, 1]. \quad (4)$$

This constitutes an easily verifiable consistency criterion (Hose and Hanss, 2019).

The possibility distribution satisfying this criterion is not unique. Beyond consistency, the choice of a possibility distribution to model the knowledge at hand is driven by the principle of maximum specificity (Dubois et al., 2004). If π_1 and π_2 are two possibility distributions such that $\pi_1(x) \leq \pi_2(x) \forall x \in \mathcal{X}$, then π_1 is said more specific than π_2 and is more informative (i.e. less conservative). Maximum specificity w.r.t. the probabilistic information (*a priori* unknown) is achieved when the possibility distribution is probabilistically calibrated¹:

$$P(C_\pi^\alpha) = 1 - \alpha, \forall \alpha \in [0, 1]. \quad (5)$$

This means that each α -cut represents a frequentist confidence interval at level $1 - \alpha$ for the variable of interest and π is a consonant confidence structure (Balch, 2020).

By construction, the individual possibility distributions $\pi(x_t | \tilde{x}_t^m \in b_j)$ verify Equation (1) with a guaranteed confidence level β . π_{EPS} being made of their envelope, it cannot be more specific than any single one of them and consequently the same guarantee applies. In the case of its α -cuts, this reads:

$$P\left(P(x_t \in C_\pi^\alpha) \geq 1 - \alpha\right) \geq \beta. \quad (6)$$

(Masson and Denceux, 2006) show empirically that their data-to-possibility transformation is rather conservative and provides a possibility distribution that actually dominates the true probability distribution with a rate much higher than the guaranteed β . Even for small sample sizes, the choice of β is not critical and quasi perfect coverage rate is obtained: $\beta \geq 0.8$, ensures that $P\left(P(x \in C_\pi^\alpha) \geq 1 - \alpha\right) \rightarrow 1$. Under this assumption, the $(1 - \alpha)$ -cuts can be used as candidate confidence intervals of *guaranteed* level α . Ideally, we are looking for $(1 - \alpha)$ -cuts verifying Equation (5), which ensures optimal specificity of π_{EPS} and thus maximally informative confidence intervals.

4. Experiments

4.1. EXPERIMENTAL SETTING

We reproduce the experiment designed by (Williams et al., 2014), who used an imperfect L96 model to investigate the performances of ensemble postprocessing for the prediction of extreme events. The system dynamics is governed by the following system of coupled equations, where the

¹ Indeed, any conservative statement such as $\exists \gamma \mid \pi(x) \geq \gamma, \forall x$ implies that $P(C_\pi^\alpha) = 1 \forall \alpha \leq \gamma$. Equation (5) ensures that a possibility distribution showing such conservative properties is discarded when compared to a possibility distribution that does not show them.

Noémie Le Carrer

X variables represent slow-moving, large-scale processes, while Y variables represent small-scale, possibly unresolved, physical processes:

$$\frac{dX_j}{dt} = X_{j-1}(X_{j+1} - X_{j-2}) - X_j + F - \frac{hc}{b} \sum_{k=1}^K Y_{j,k} \quad (7)$$

$$\frac{dY_{j,k}}{dt} = cbY_{j,k+1}(Y_{j,k-1} - Y_{j,k+2}) - cY_{j,k} + \frac{hc}{b} X_j \quad (8)$$

where $j = 1, \dots, J$ and $k = 1, \dots, K$. The parameters are set to: $J = 8$, $K = 32$, $h = 1$, $b = 10$, $c = 10$ and $F = 20$. This perfect model is randomly initialised and then integrated forward in time by means of a Runge-Kutta 4th-order method with time step $dt = 0.002$ (model time units) until enough trajectories of duration 1.4, starting every 1.5 time units, are recorded for our analysis. A lead time $t = 1$ corresponds to 0.2 model time units after initialisation and can be associated with approximately 1 day in the real world (Lorenz, 1996). We are interested in predicting the variable X_1 .

An imperfect version of the L96 system is implemented to generate predictions for the X_j variables. In Equation (7), $-\frac{hc}{b} \sum_{k=1}^K Y_{j,k}$ is replaced with:

$$0.262 - 1.262X_j + 0.004608X_j^2 + 0.007496X_j^3 - 0.0003226X_j^4 \quad (9)$$

To reproduce the perturbation of the ICs, M perturbed members $\tilde{X}_j$ are sampled independently around the true value of each variable X_j following a normal distribution $\tilde{X}_j \sim \mathcal{N}(X_j, 0.1^2)$. These ensemble sets are initialised each time a new trajectory record starts, and integrated forward in time up to lead time 1.4 by means of a Runge-Kutta 4th-order method with lower time resolution ($\tilde{dt} = 0.02$ model time units). The size of the ensemble is set to $M = 24$, a value comparable to operational weather forecasting schemes (e.g. $M = 17$ for the Met Office Global and Regional Ensemble Prediction System).

4.2. REFERENCE MODEL: GAUSSIAN ENSEMBLE DRESSING

We compare the performances of our approach (POSS hereafter) to those of a classical probabilistic framework for interpreting EPSs, namely a Gaussian ensemble dressing (GEB hereafter). Its predictive probability distribution reads (Roulston and Smith, 2003):

$$p(x_t|\tilde{\mathbf{x}}_t)_\theta = \frac{1}{M} \sum_{i=1}^M \mathcal{N}(a\tilde{x}_t^i + \omega, \sigma^2) \quad (10)$$

We infer the parameters $\theta = \{a, \omega, \sigma\}$ through the optimisation of the ignorance score (Roulston and Smith, 2002) over the archive $\mathcal{I}_t$ used in the possibilistic framework. To that end, we use the nonlinear programming solver provided by MATLAB® and apply the guidance developed in (Bröcker and Smith, 2008) to provide robust solutions.

Confidence intervals at level α on x_t are obtained from p by a method that provides the desired intervals associated with the highest-density regions (Hyndman, 1996). We also report in the next

A possibilistic interpretation of ensemble predictions

section the performances of the confidence intervals similarly extracted from the unprocessed probability density (hereafter RAW) associated with the EPS (a histogram of the EPS normalised to represent a probability density).

4.3. EVALUATION CRITERIA

We aim at answering the questions:

- (a) Can a possibilistic treatment of the EPS provide more guarantees than a probabilistic interpretation?
- (b) If yes, at what cost?

To that end, we compare the performances of the confidence intervals at level α , noted I^α , extracted from the methodologies POSS, GEB and RAW as described in the previous sections. We say that a confidence interval is *guaranteed at level α* if the coverage probability verifies $P(x \in I^\alpha) \geq \alpha$. We use the term *guaranteed* in the sense that such an interval is associated with a lower bound on the (frequentist) probability that the verification falls within it. Such guarantees are sought e.g. in risk-averse decision-making. We say that it is *reliable*, or *probabilistically calibrated*, when $P(x \in I^\alpha) \approx \alpha$. We call it all the more *conservative* than $P(x \in I^\alpha) - \alpha$ is large, which is associated with non optimal interval *precision*.

4.4. EXPERIMENTS

All results presented here use $n = 30$ bins of similar width to partition the x -axis². The test set consists in 40,000 independent trajectories of length $t = 7$ days and the corresponding EPS predictions. All EPSs have beforehand been preprocessed to remove the constant bias. We consider a range of archive size $N_I \in \{156, 1560, 5 \times 10^3, 15 \times 10^3, 30 \times 10^3\}$. In particular, $N_I = 156$ corresponds to 3 years of model archive, whereas $N_I = 1560$ amounts to 30 years, which corresponds to the standard length of a historical re-forecast dataset (Hamill et al., 2004; Hagedorn, 2008). The two latter N_I are operational figures, unlike larger values that we present to study the asymptotic properties of our framework.

We define two types of events: an extreme event, " $x \leq q_5$ " (EE), and a common event, " $q_{50} < x \leq q_{55}$ " (NEE) where q_i represents the percentile of level i of the climatic distribution of x (i.e. global distribution), plotted in Figure 4 along with both events. This will allow us to use test sets of similar sizes³ in order to position our approach against the generic probabilistic postprocessing techniques that are known to weakly address such extreme events.

A preliminary assessment (Figure 5) of the effect of the parameter β of Goodman's model on the probabilistic reliability of the $(1 - \alpha)$ -cuts derived from π_{EPS} shows that varying β from 0.6 to 1 does not impact guarantees at any given N_I for the events of interest. It only impacts precision

² This choice is based on the range covered by the climatology of x and the fact that x can be associated to a physical quantity of the atmosphere, e.g. temperature, which leads to bins of width ≈ 2 degrees. For other systems and applications, the bins can be for instance partitioned so that the distribution of the climatology is homogeneous over the bins.

³ About 2×10^3 elements.

Noémie Le Carrer

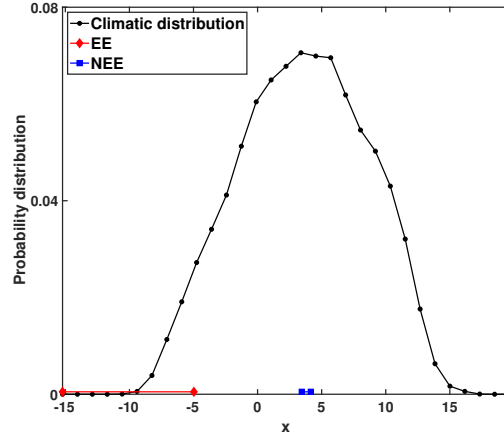


Figure 4. Climatic distribution of the L96's variable of interest X_1 (x for simplicity) where the 'extreme' event " $x \leq q_5$ " (EE) and 'common' event " $q_{50} < x \leq q_{55}$ " (NEE) are reported.

and its effect is only visible for small archives ($N_I \leq 156$) or large lead times, especially in the EE case. We consequently use $\beta = 0.9$ in our experiments, which allows to improve specificity while maintaining guarantees on confidence intervals.

5. Results

5.1. EMPIRICAL ASSESSMENT OF FORMAL GUARANTEES

Figure 6 reports the coverage probability of the confidence intervals I^α extracted for $\alpha \in \{0, 0.05, 0.1, \dots, 1\}$ for all evaluated methodologies at lead times $t \in \{1, 3, 5, 7\}$ days. We first note that using RAW leads to confidence intervals that are not guaranteed for $t > 1$ day for both EE and NEE. Postprocessing (here GEB) allows to make them guaranteed at all lead times for the NEE and for $t \leq 3$ days for the EE. The effect of the training set size for the probabilistic treatment does not appear to be significant. Conversely, the confidence intervals derived using POSS are globally guaranteed for both events and at all lead times for operational archives ($N_I < 5 \times 10^3$). Interestingly, when the archive grows significantly, confidence intervals with large α are not guaranteed anymore for the larger lead times in the EE case. The effect appears all the earlier (in terms of lead time) than N_I is large.

We observe here a limitation of possibility theory: its strength lies in incomplete information. As shown in Figure 2, the larger the datasets used to derive possibility distributions, the closer the possibility distribution is in shape to the underlying probability distribution. In particular, the level γ such as $\pi(x) \geq \gamma \forall x$ tends towards zero. In other words, such possibility distributions tend to conceal the possibility of rare events.

We illustrate this phenomenon in Figure 7, where we represent the average density of analogs used to compute the individual $\pi(x_t | \tilde{x}_t^m \in b_j)$ (see step 3, Figure 3). In the EE case, as the lead time increases, this average density decreases by several orders of magnitude for the more extreme

A possibilistic interpretation of ensemble predictions

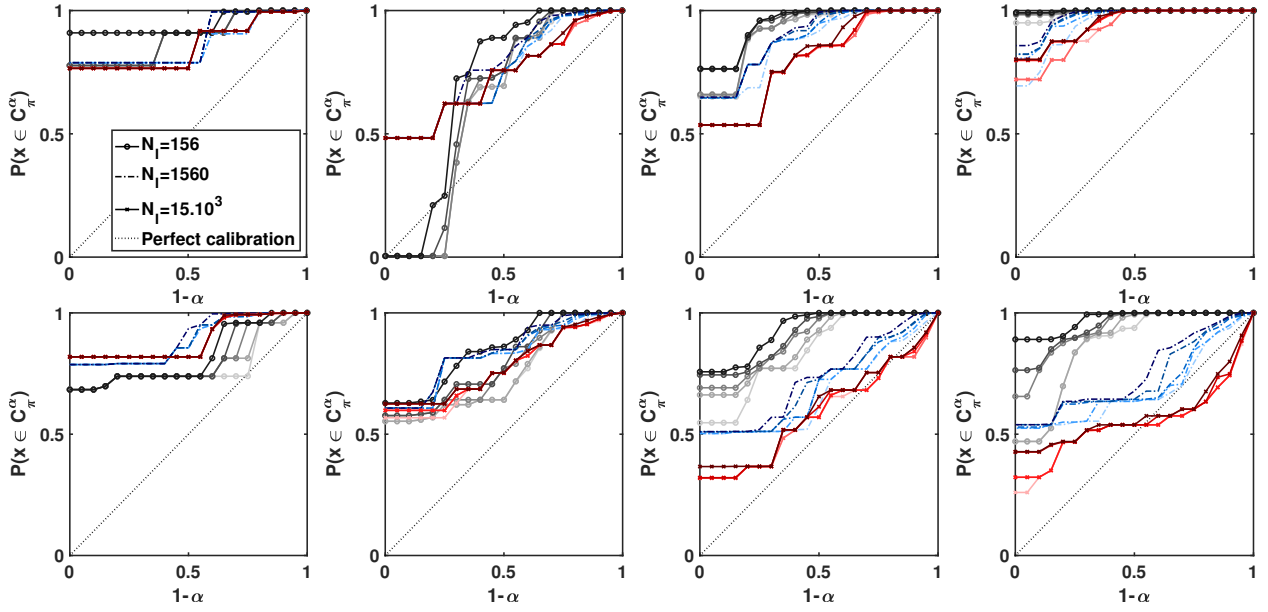


Figure 5. Coverage probability of the α -cuts of π_{EPS} at lead time $t \in \{1, 3, 5, 7\}$ days (left to right), in the case of the NEE (top) and EE (bottom). Goodman models with parameter $\beta \in \{0.6, 0.75, 0.9, 0.95, 0.99\}$ (the darker the line, the larger β) are compared in the case of three archives of respective size $N_I \in \{156, 1560, 15 \times 10^3\}$ (grey, blue and red color scales respectively).

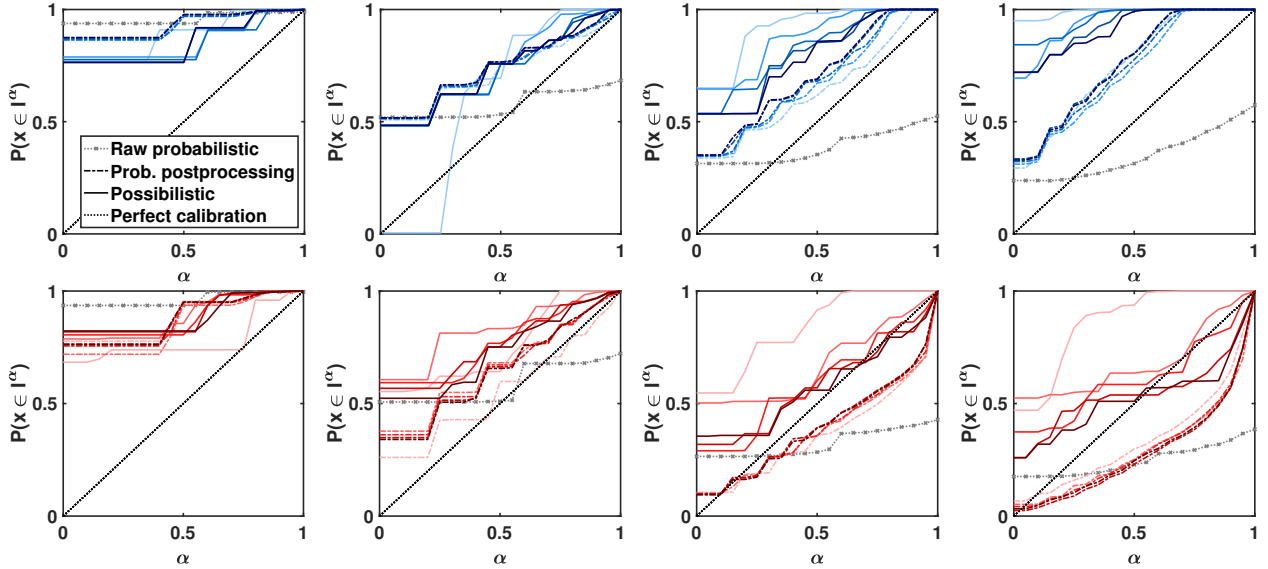


Figure 6. Coverage probability of the $(1 - \alpha)$ -cuts of π_{EPS} used as confidence intervals of level α at lead time $t \in \{1, 3, 5, 7\}$ days (left to right), in the case of the NEE (top) and EE (bottom). The EPS archive size is $N_I \in \{156, 1560, 5 \times 10^3, 15 \times 10^3, 30 \times 10^3\}$ (the larger the darker the line). The coverage probability of the confidence intervals of level α derived from the raw EPS's probability density and from the postprocessed density (with the same training set of size N_I as used in the possibilistic framework) is reported as well.

Noémie Le Carrer

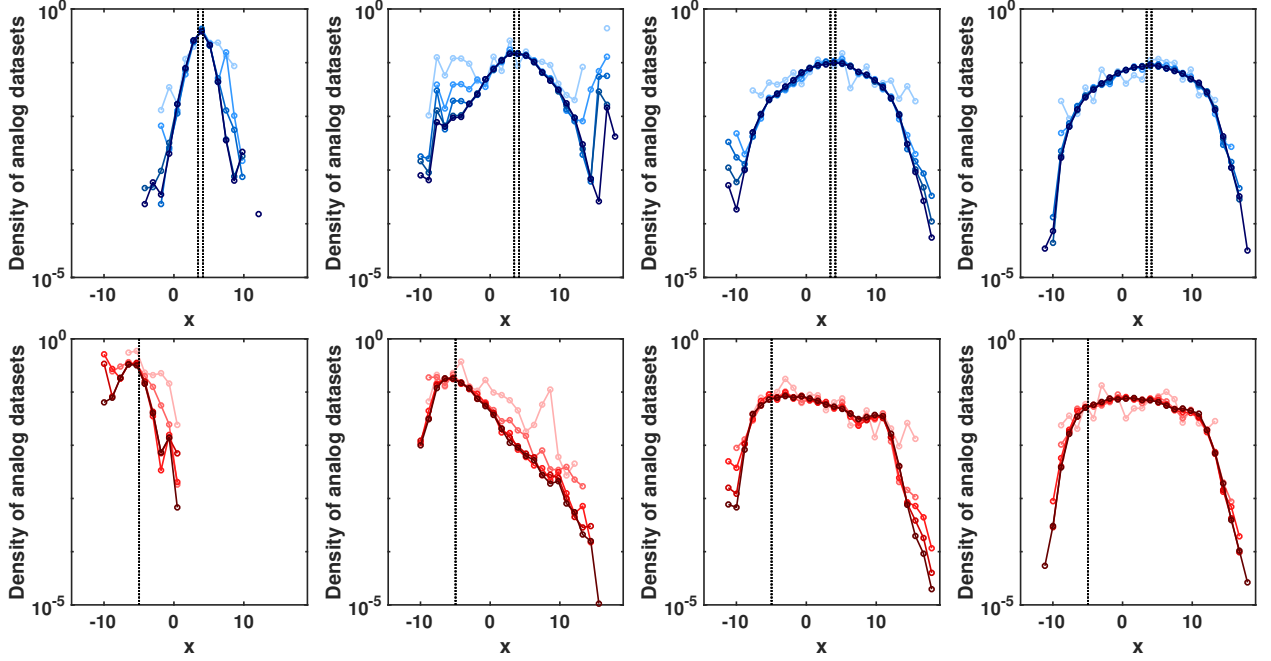


Figure 7. Average density of the analog datasets used to derive π_{EPS} , for sizes $N_I \in \{156, 1560, 5 \times 10^3, 15 \times 10^3, 30 \times 10^3\}$ (the larger N_I the darker the line) and lead time $t \in \{1, 3, 5, 7\}$ days (left to right), in the case of the NEE (top) and the EE (bottom). Only densities above 0 are represented. Vertical dotted lines allow to visualise the events of interest (note that the EE is only defined by its upper bound).

bins ($x \rightarrow \inf \mathcal{X}$). This drop is all the more significant than N_I is large. For small $N_I \leq 1560$, the more extreme bins are, as expected, not represented but the intermediary bins are and their density remains above $\frac{1}{100}$. For very large $N_I \geq 5 \times 10^3$, the more extreme bins are represented however their density drops below $\frac{1}{1000}$. In other words, the rarest events part of EE are represented only for extremely large archives, where they will be part of large analog sets, which implies, given the asymptotic behaviour illustrated in Figure 2, that they will be concealed from the associated possibility distributions. More precisely, the level γ such as $\pi_{EPS}(x) \geq \gamma \quad \forall x \in \mathcal{X}$ remains strictly positive so $P(x \in I^\alpha) = 1$ remains valid for $\alpha \approx 1$ (that is the large scale $(1 - \alpha)$ -cuts where $1 - \alpha \rightarrow 0$). However for intermediate α , the $(1 - \alpha)$ -cuts may not extend enough towards extreme bins, which negatively impacts the coverage rate. This trend is only observed for sufficiently large α , as possibility distributions remain globally more conservative than the EPS-based probability distributions (see next Section), and consequently provide I^α that encompass more observations than the frequentist calibration requires in the case of smaller α (i.e. for the upper part of the distribution). The "sufficiently large α " decreases with increasing lead times and archive sizes, following the effect described in Figure 7. Figure 8 illustrates our point by breaking down the coverage probability for three subsets of the EE: large archives lead to POSS-based confidence intervals that are all the more guaranteed as the event of interest is not too extreme. Probabilistic calibration for the more extreme part of EE can be improved by increasing the parameter β , however this has no effect in the case of large archives (see Figure 9).

A possibilistic interpretation of ensemble predictions

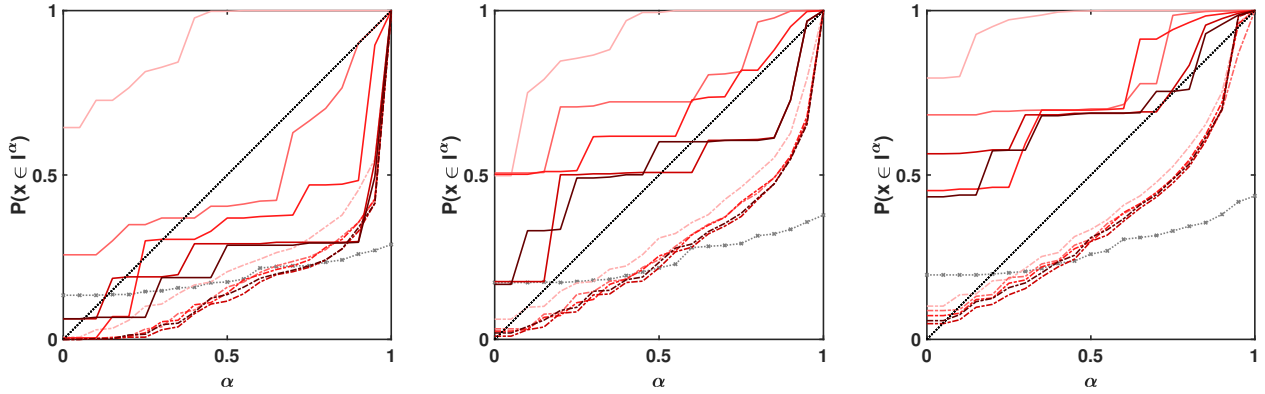


Figure 8. Coverage probability of the $(1 - \alpha)$ -cuts of π_{EPS} at lead time $t = 7$ days, in the case of events belonging to a partition of subsets of EE (from left to right: $x \leq q_1$, $q_1 < x \leq q_3$ and $q_3 < x \leq q_5$). The EPS archive size varies: $N_I \in \{156, 1560, 5 \times 10^3, 15 \times 10^3, 30 \times 10^3\}$ (the larger the darker the line). The probabilistic calibration of the confidence intervals of level α derived from the raw EPS's probability density and from the postprocessed density (with the same training set of size N_I as the possibilistic framework) is also reported. See Figure 6 for legend.

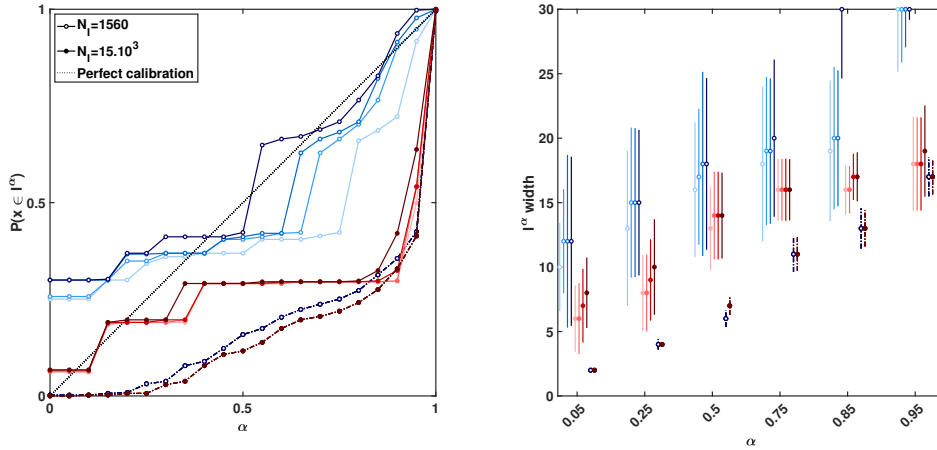


Figure 9. Coverage probability (left) and associated width distributions (right) of the confidence intervals of level α at lead time $t = 7$ days, in the EE case, for two archive sizes $N_I \in \{1560, 15 \times 10^3\}$ (blue and red color scale respectively). POSS results (solid line) for increasing Goodman's parameter $\beta \in \{0.6, 0.9, 0.95, 0.99\}$ (the larger the darker the line) are compared to GEB results (dotted line). The width distribution is represented through its mean and one standard deviation above and below.

The NEE case study does not suffer from this limitation as the density of analogs falling in the NEE bins remains around $\frac{1}{10}$ at all lead times. In comparison to GEB, POSS improves the reliability of confidence intervals for very short lead times while they remain more conservative for large lead times.

Provided that N_I is not too large (which we assume is always the case for operational archives), Figures 6 and 8 clearly show the added value of treating the EPS in a possibilistic manner in terms of guarantees for the EE at large lead times, or in terms of reliability for the NEE at very small lead

Noémie Le Carrer

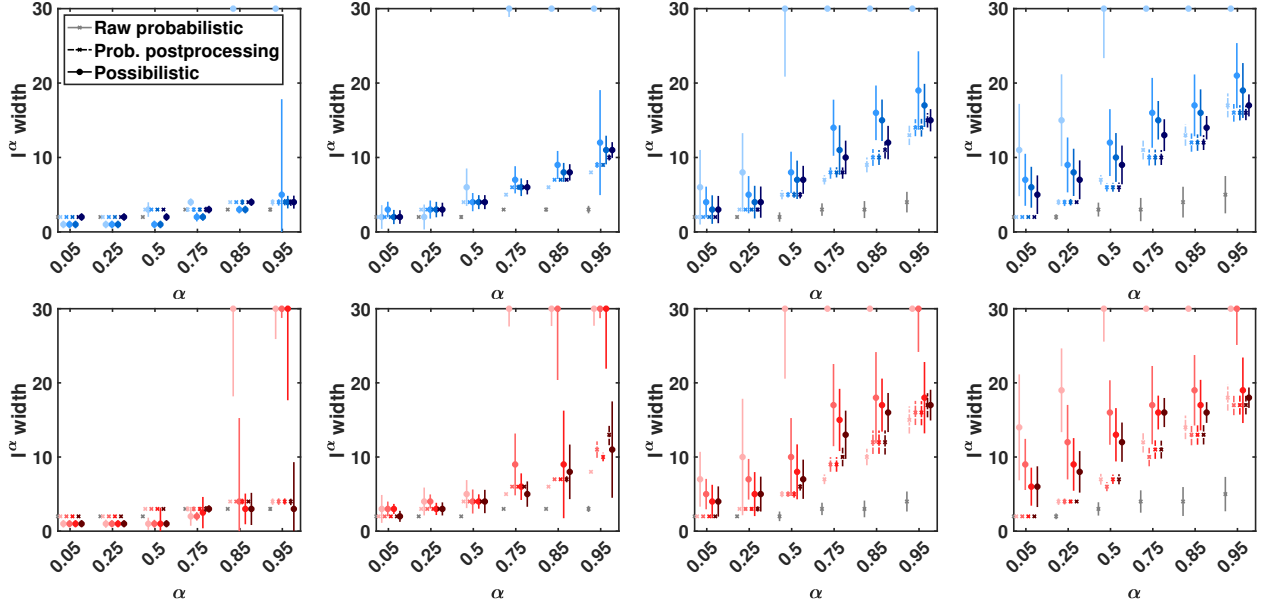


Figure 10. Distribution (mean $\pm$ standard deviation) of the width of the possibility and probability-based confidence intervals described in the legend of Figure 6 for lead time $t \in \{1, 3, 5, 7\}$ days (left to right), in the case of the NEE (top) and EE (bottom). Only the cases $N_I \in \{156, 1560, 5 \times 10^3, 30 \times 10^3\}$ are represented (the larger N_I , the darker the line).

times. However, we can wonder what is the cost of such improvements. How do the possibility-based confidence intervals compare to their probability-based counterparts, in terms of precision?

5.2. INTERVAL PRECISION

Figure 10 compares the average width of the confidence intervals derived from the three methodologies. For both EE and NEE, N_I affects the width of the possibilistic I^α significantly, making them narrower with larger N_I , all the more than the lead time increases. Their probabilistic counterparts are generally much smaller, except when $N_I \approx 30 \times 10^3$.

For NEE and level $\alpha < 0.9$, POSS brings more information at very short lead times ($t = 1$ day) than the probabilistic approaches: intervals are smaller or equal in size and remain guaranteed. This is all the more true that the archive is of intermediate size ($N_I = 1560$). Increasing the lead time beyond $t = 3$ days favors the probabilistic approach, which is more reliable with narrower intervals.

For EE, the added value of POSS over GEB is observed on two occasions: 1) intervals are as reliable yet narrower for very small lead times and $\alpha < 0.9$, whatever the archive size ; 2) for large lead times and intermediary-sized archives ($N_I \in \{1560, 5 \times 10^3\}$), possibility-based confidence intervals are both guaranteed, reliable and operational (i.e. not too wide compared to GEB's results, contrary to what $N_I = 156$ produces), while the probabilistic intervals are narrower yet not guaranteed at all. In the case of particularly rare events, as represented in Figure 8, an intermediary archive such as $N_I = 1560$ is able to produce confidence intervals close to perfect reliability even

A possibilistic interpretation of ensemble predictions

for large lead times, as long as the parameter β is increased towards 1. Such reliability is reached at the expense of the interval width, which is significantly increased (w.r.t. smaller β) for the largest $\alpha \geq 0.85$.

6. Conclusion

We introduced a novel framework to interpret EPSs where a possibility distribution π_{EPS} is derived from the EPS at hand and an archive of (EPS; verification). We showed how to use the $(1-\alpha)$ -cuts of a continuous interpretation of π_{EPS} to produce confidence intervals at level α about the future value of the variable of interest. Our possibility-based confidence intervals come with formal guarantees, and experimental results show that they overpass probability-based ones in two situations: 1) at very small lead times for both common and extreme events, where they are as reliable yet narrower; 2) more blatantly, at intermediate and large lead times for extreme events, where they remain guaranteed and can be brought close to perfect reliability even for particularly rare events, yet at the expense of precision. These results can be reached with operational archive like the 20 – 30-year reforecast datasets. The guarantees are retained for smaller archives, which however lead to more conservative intervals and thereby impede operationality.

Possibility theory is a promising tool for the prediction of extreme events, given a limited and imperfect amount of information on the system's dynamics. Beyond the results presented in this article, further developments by the author (Le Carrer and Ferson, 2020) show how π_{EPS} can be combined with additional possibility distributions constructed from alternative sources of information such as the IC or dynamical information (see step 6 of Figure 3). Therein, the concept of ignorance briefly introduced in Section 2 is developed and presented as a promising tool for risk communication.

Acknowledgements

This research is funded by the Engineering & Physical Sciences Research Council (EPSRC) and the Economic & Social Research Council (ESRC) with grant no. EP/L015927/1. We are very grateful to N. Berthier for his careful reviewing.

References

- Balch, M. S.: 2020, 'New Two-Sided Confidence Intervals for Binomial Inference Derived Using Walley's Imprecise Posterior Likelihood as a Test Statistic'. *International Journal of Approximate Reasoning*.
- Bröcker, J. and L. A. Smith: 2008, 'From ensemble forecasts to predictive distribution functions'. *Tellus A: Dynamic Meteorology and Oceanography* **60**(4), 663–678.
- Buizza, R.: 2018, 'Ensemble Forecasting and the Need for Calibration'. In: *Statistical Postprocessing of Ensemble Forecasts*. Elsevier, pp. 15–48.
- Cayrac, D., D. Dubois, M. Haziza, and H. Prade: 1994, 'Possibility theory in "Fault mode effect analyses". A satellite fault diagnosis application'. In: *Proceedings of 1994 IEEE 3rd International Fuzzy Systems Conference*. pp. 1176–1181.

Noémie Le Carrer

- Dempster, A. P.: 2008, 'Upper and Lower Probabilities Induced by a Multivalued Mapping'. In: R. R. Yager and L. Liu (eds.): *Classic Works of the Dempster-Shafer Theory of Belief Functions*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 57–72.
- Dubois, D., L. Foulloy, G. Mauris, and H. Prade: 2004, 'Probability-Possibility Transformations, Triangular Fuzzy Sets, and Probabilistic Inequalities'. *Reliable computing* **10**(4), 273–297.
- Dubois, D. and H. Prade: 2012, *Possibility theory: an approach to computerized processing of uncertainty*. Springer Science & Business Media.
- Friederichs, P. and A. Hense: 2007, 'Statistical Downscaling of Extreme Precipitation Events Using Censored Quantile Regression'. *Monthly Weather Review* **135**(6), 2365–2378.
- Friederichs, P., S. Wahl, and S. Buschow: 2018, 'Postprocessing for Extreme Events'. In: S. Vannitsem, D. S. Wilks, and J. W. Messner (eds.): *Statistical Postprocessing of Ensemble Forecasts*. Elsevier, pp. 127–154.
- Gneiting, T. and M. Katzfuss: 2014, 'Probabilistic forecasting'. *Annual Review of Statistics and Its Application* **1**, 125–151.
- Goodman, L. A.: 1965, 'On simultaneous confidence intervals for multinomial proportions'. *Technometrics* **7**(2), 247–254.
- Hagedorn, R.: 2008, 'Using the ECMWF reforecast dataset to calibrate EPS forecasts'. *ECMWF Newsletter* **117**, 8–13.
- Hamill, T. M., J. S. Whitaker, and X. Wei: 2004, 'Ensemble reforecasting: Improving medium-range forecast skill using retrospective forecasts'. *Monthly Weather Review* **132**(6), 1434–1447.
- Hose, D. and M. Hanss: 2019, 'Possibilistic calculus as a conservative counterpart to probabilistic calculus'. *Mechanical Systems and Signal Processing* **133**, 106290.
- Hyndman, R. J.: 1996, 'Computing and graphing highest density regions'. *The American Statistician* **50**(2), 120–126.
- Le Carrer, N. and S. Ferson: 2020, 'Beyond probabilities: A possibilistic framework to interpret ensemble predictions and fuse imperfect sources of information'. Under review.
- Legg, T. P. and K. R. Mylne: 2004, 'Early Warnings of Severe Weather from Ensemble Forecast Information'. *Weather and Forecasting* **19**(5), 891–906.
- Lorenz, E. N.: 1996, 'Predictability: A problem partly solved'. In: *Proc. Seminar on predictability*, Vol. 1.
- Masson, M.-H. and T. Dencoux: 2006, 'Inferring a possibility distribution from empirical data'. *Fuzzy Sets and Systems* **157**(3), 319–340.
- Mylne, K., C. Woolcock, J. Denholm-Price, and R. Darvell: 2002, 'Operational calibrated probability forecasts from the ECMWF ensemble prediction system: implementation and verification'. In: *Preprints of the Symposium on Observations, Data Assimilation and Probabilistic Prediction*. pp. 113–118.
- Roulston, M. S. and L. A. Smith: 2002, 'Evaluating Probabilistic Forecasts Using Information Theory'. *Monthly Weather Review* **130**(6), 1653–1660.
- Roulston, M. S. and L. A. Smith: 2003, 'Combining dynamical and statistical ensembles'. *Tellus A: Dynamic Meteorology and Oceanography* **55**(1), 16–30.
- Smith, L. A.: 2016, 'Integrating Information, Misinformation and Desire: Improved Weather-Risk Management for the Energy Sector'. In: P. J. Aston, A. J. Mulholland, and K. M. Tant (eds.): *UK Success Stories in Industrial Mathematics*. Cham: Springer International Publishing, pp. 289–296.
- Williams, R. M., C. A. T. Ferro, and F. Kwasniok: 2014, 'A comparison of ensemble post-processing methods for extreme events'. *Quarterly Journal of the Royal Meteorological Society* **140**(680), 1112–1120.
- Zadeh, L.: 1978, 'Fuzzy sets as a basis for a theory of possibility'. *Fuzzy Sets and Systems* **1**(1), 3–28.