Beyond probabilities: A possibilistic framework to interpret ensemble predictions and fuse imperfect sources of information

Noémie Le Carrer[1,3] and Scott Ferson[2]

[1,2] Institute for Risk and Uncertainty, Department of Engineering, University of Liverpool, United Kingdom
[3] Dipartimento di Scienze Ambientali, Informatica & Statistica, Università Ca'Foscari Venezia, Italy

Correspondence: N. Le Carrer (noemie.lecarrer@unive.it)

Abstract. Ensemble forecasting is widely used in medium-range weather predictions to account for the uncertainty that is inherent to the numerical prediction of high-dimensional, nonlinear systems with high sensitivity to initial conditions. Ensemble forecasting allows one to sample possible future scenarii in a Monte-Carlo-like approximation through small strategical perturbations of the initial conditions, and in some cases stochastic parameterisation schemes of the atmosphere-ocean dynamical equations. Results are generally interpreted in a probabilistic manner by turning the ensemble into a predictive probability distribution. Yet, due to model bias and dispersion errors, this interpretation is often not reliable and statistical postprocessing is needed to reach probabilistic calibration. This is all the more true for extreme events that for dynamical reasons, cannot generally be associated with a significant density of ensemble members.

In this work we propose a novel approach: a possibilistic interpretation of ensemble predictions, taking inspiration from possibility theory. This framework allows us to integrate in a consistent manner other imperfect sources of information, such as the insight about the system dynamics provided by the analog method. We thereby show that probability distributions may not be the best way to extract the valuable information contained in ensemble prediction systems, especially for large lead times. Indeed, shifting to possibility theory provides more meaningful results without the need to resort to additional calibration, while maintaining or improving skills. Our approach is tested on an imperfect version of the Lorenz 96 model, and results for extreme event prediction are compared against those given by a standard probabilistic ensemble dressing.

Key-words: *Ensemble prediction, Probabilistic weather forecasting, Recalibration, Statistical post-processing, Extreme event, Weather regimes, Possibility theory, Imprecise probabilities*

1 Introduction

Predicting the weather through numerical models of the atmosphere is impeded by the mere nature of the atmospheric dynamics, characterised by strong nonlinearities and high sensitivity to initial conditions. Limited grid resolution in the initial

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/qj.4135

conditions (ICs), discrepancies introduced by measurement errors and incomplete description of the system's dynamics, contribute to error growth and limit the skill of short and medium-range point predictions. A shift in paradigm was introduced in parallel to the increase of computational resources at the beginning of this century, when low-resolution ensemble predictions started to replace, or complete, the traditional single high-resolution deterministic prediction. The idea behind these ensemble forecasts had been developed earlier by Leith (1974), who suggested to sample M ICs around the actual best ICs estimation, to run the model forward for each IC, and to interpret the M resulting predictions in a Monte-Carlo like fashion. Ensemble forecasts are thus interpreted in a probabilistic way, either to characterise the predictability of the associated deterministic forecast (e.g. through the variance of the ensemble) or to directly provide probabilities of observing a given event.

Probabilistic interpretation of ensemble predictions

However, such a probabilistic interpretation poses conceptual issues. First, the ICs are perturbed according to schemes designed to sample in a minimalist way particularly high-dimensional systems like numerical weather global models. These schemes generally select the initial perturbations leading to the fastest growing perturbations (e.g. singular vectors (Hartmann et al., 1995), bred vectors (Toth and Kalnay, 1997)). Although this way of proceeding is an efficient manner to detect the range of possible futures, one cannot consider that the M perturbed ICs are random samples, and consequently cannot interpret the resulting ensemble as a sample of the distribution characterising the future state of the system. Besides, one of the core assumptions of Leith (1974) is that model error is negligible w.r.t. the error resulting from the propagation of the uncertainty on the ICs. In practice, the assumption of such near-perfect models is not always true and after a few hours, the convex hull of the ensemble trajectories is not guaranteed to contain the observed trajectory, traducing structural bias (Toth and Kalnay, 1997; Orrell, 2005).

The above conceptual issues impede a probabilistic interpretation of ensembles prediction systems (EPSs) in practice: despite the introduction of stochastic parameterisation schemes to account for model error (Buizza et al., 1999), the operational ensembles remain overconfident, i.e. with a spread that is generally too small (Wilks and Hamill, 1995; Buizza, 2018). In particular, the predictive probabilities derived from ensemble forecasts are not reliable. On average, the probability derived for a given event does not equal the frequency of verification (Bröcker and Smith, 2007; Hamill and Scheuerer, 2018). Although such probabilistic predictions have higher forecast skill than the climatology, most often they cannot be used as actionable probabilities. By design (limited EPS size, targeted sampling of ICs) and by context (flow-dependent regime error, strongly nonlinear system) they do not represent the true probabilities of the system at hand (Legg and Mylne, 2004; Bröcker and Smith, 2008). This verification is all the more true for extreme events, that result from nonlinear interactions at every and between

scales. Such interactions cannot be reproduced in number in a limited-size ensemble prediction system (Legg and Mylne, 2004), which implies that extreme events generally cannot be associated to a high density of ensemble members.

Biases and dispersion errors in ensemble forecasts consequently call for statistical postprocessing to improve the information content and calibration of probabilistic predictions (Gneiting and Katzfuss, 2014; Buizza, 2018). A range of methods have been developed to address the above-mentioned limitations. The most classical ones fit an optimised parametric distribution either: a) onto each ensemble member, and aggregate them all to provide a global probability density function (PDF) (e.g. Bayesian model averaging, introduced by Raftery et al. (2005)); or b) onto the whole ensemble, with parameters derived from linear combinations of the ensemble's characteristics (non-homogeneous regression, developed by Gneiting et al. (2005)). More specific approaches target for instance the improvement of reliability, e.g. rank histogram recalibration (Hamill and Colucci, 1997) which makes use of the information content of the rank histogram to issue ensemble-based predictions that show better probabilistic calibration. More recently, calibration by means of the probability integral transform was suggested by Graziani et al. (2019), while Smith (2016) developed a user-oriented framework based on the actual probability of success for a given probabilistic threshold, and Hamill and Scheuerer (2018) developed a framework based on quantile mapping and rank-weighted best-member dressing over single or multimodel EPSs.

Although generic postprocessing strategies do improve the predictive skill for common events, they tend to deteriorate the results for extreme events (Mylne et al., 2002), which consequently need separate and tailored treatment. Friederichs et al. (2018) shows that when the tail of the climatology is short, a flexible skewed distribution (e.g. a generalised extreme value distribution as suggested by Scheuerer (2014)) for the complete sample space is a good solution for predicting extremes as well. However, a separate description of the tail distribution by means of quantile regression (Friederichs and Hense, 2007) or nonstationary Poisson process (Friederichs et al., 2018) may be necessary in the case of heavy climatology tails.

Possibility theory and EPSs

In view of all this, and especially considering the need to resort to (possibly multiple) calibration steps to provide meaningful probabilistic outputs, we echo Bröcker and Smith (2008) who question the choice of probability distributions as *the best representation of the valuable information contained in an EPS*. Rather, we wonder whether possibility theory, "a weaker theory than probability [...] also relevant in non-probabilistic settings where additivity no longer makes sense" (Dubois et al., 2004), provides an interesting alternative, in a context where conceptual and practical limitations restrict the applicability of a density-based (i.e. additive) interpretation of EPSs.

This is what we investigate in this work. We have shown in a previous study (Le Carrer and Green, 2020), that using a possibilistic ensemble dressing to calibrate the predictive probabilities instead of its probabilistic counterpart incurred two important limitations: 1) its parametric form introduced trade-off in performances as well as the impossibility to propagate the formal guarantees that possibility theory provides, and 2) the local dynamics of the system was not explicitly taken into account. In this article, we go further and address these two main limitations.

Regarding point 2), just like a global probabilistic interpretation of EPSs misses the introduction of state-dependent refinement that allows parameters to adapt to different regimes of model error (Orrell, 2005; Allen et al., 2019), a purely ensemble-based framework may be too conservative due to a lack of information about the dynamics of the system (noted $\mathcal{S}$ hereafter) at the time of interest. We consequently combine our possibilistic interpretation of EPSs to a method providing dynamical analogs, in our case the empirical dynamic modeling of $\mathcal{S}$. The underlying assumption of resorting to analogs is the existence of a deterministic structure governing the co-evolution of the coupled variables of $\mathcal{S}$. The underlying structure of such a system is revealed by the state dependent dynamics occurring on a strange attractor manifold $\mathcal{A}$. Takens' delay embedding theorem (Takens, 1981) and its generalisation by Deyle and Sugihara (2011), describe how lagged variables of a single time series, or combinations of several coupled time series, can be used to reconstruct a shadow attractor $\mathcal{A}'$ of $\mathcal{A}$, that is a smooth and smoothly invertible 1:1 mapping with $\mathcal{A}$. Making predictions from the shadow attractor consists in finding the closest neighbors of the ICs of interest in the attractor, following their trajectories up to the desired lead time, and retrieving the corresponding so-called analog predictions. These are then used to construct, e.g. a probabilistic prediction for the target day. In practice, finding true analogs in a time series for high-dimensional systems such as the atmosphere-ocean is a difficult task (Lorenz, 1969; Van den Dool, 1994). Similarity-based methods (also coined as analog methods) were developed, applying the same philosophy yet on a reduced number of variables characterising the system, that is without taking into account its full dimensionality. Thus statistical downscaling, based on the hypothesis that two close synoptic situations may produce close local effects (Lorenz, 1956, 1969), is used for operational precipitation forecasting (Hamill and Whitaker, 2006; Daoud et al., 2016). Common analog forecasting operators are presented in Platzer et al. (2021) and their respective properties and performances are analysed from a theoretical point of view, connecting analog forecasting error to local approximations of the system's dynamics. Empirical dynamical modelling, locating analogs in the shadow attractor space or in one of its sub-spaces, is still used to perform model-free predictions (Ma et al., 2017) or to give insight on predictability (Trevisan, 1995; Ramesh and Cane, 2019).

Generally speaking, making predictions from analogs performs all the more as the record of one or more variable(s) describing $\mathcal{S}$ is long, and as $\mathcal{S}$ is of small dimension. Still, we posit that using possibility theory to interpret analogs allows us

to extract more dynamical information from the incomplete shadow attractor reconstruction than a PDF or a weighted mean of analogs. Besides, such a choice allows us to combine this additional source of information to the EPS information in a consistent language of reference, particularly well suited to the fusion of information.

Summary of contributions and outline

In this work, we investigate the benefits of: (i) using a framework based on possibility theory for extracting the information contained in an EPS; and (ii) combining it with the insight about the local dynamics of the system gained from the analog method. Our investigation is particularly driven by the following three questions:

- Can we draw an interpretation framework of EPS that would directly make sense and provide outputs that are meaningful without having to resort to additional layers of calibration?
- Can we simultaneously maintain or improve the prediction skills compared to those of standard probabilistic interpretations?
- Can we operationally use the possibilistic outputs at their full potential, that is more than simply deriving associated probabilities?

We support our study with numerical experiments on a commonly used surrogate model of atmospheric dynamics, namely the L96 system (Lorenz, 1996) that we present in Section 4. Section 2 introduces the basics of possibility theory, that we then use in Section 3 to develop our novel possibilistic framework for the interpretation of EPSs. Therein, we also explains how to extract and combine the dynamical information gained via the analog method. We present the modalities of assessment in Section 4. Our novel methodology is tested in the context of extreme event prediction on an imperfect version of the L96 and results are discussed in Section 5. A conclusion follows.

2 Possibility theory

2.1 Basic principles

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 π is a function $\pi : \mathcal{X} \to [0, 1]$ that represents the state of knowledge about the current state

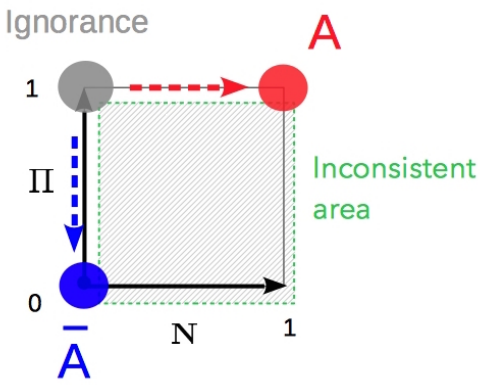


Figure 1. $N-\Pi$ diagram, depicting the dual measures of possibility theory. A is the event of interest and $\bar{A}$ its complement. The hatched area represents the area of inconsistent combinations for N and Π.

of the system. Given an event $A \subseteq \mathcal{X}$, the possibility and necessity measures are defined respectively as: $\Pi(A) = \sup_{x \in A} \pi(x)$ and $N(A) = 1 - \Pi(\bar{A})$ where $\bar{A}$ represents the complementary event of A. Π and N satisfy the following axioms:

1. $\Pi(\mathcal{X}) = 1$ and $\Pi(\varnothing) = 0$, where $\varnothing$ represents the empty set;
2. $\Pi(A \cup B) = \max\big(\Pi(A), \Pi(B)\big)$ (similar to $N(A \cap B) = \min\big(N(A), N(B)\big)$), where $B \subseteq \mathcal{X}$.

The measures can be interpreted in the following way (Dubois and Prade, 2015):

a. $N(A) = 1 \Leftrightarrow \Pi(\bar{A}) = 0$ indicates that A is necessary so it has to happen and $\bar{A}$ is impossible;

b. $0 < N(A) < 1$ is a tentative acceptance of A to a degree $N(A)$, since $\min\big(N(A), N(\bar{A})\big) = 0$ from axiom 2 ($\bar{A}$ is not necessary at all);

c. $\big(\Pi(A) = \Pi(\bar{A}) = 1\big) \Leftrightarrow \big(N(A) = N(\bar{A}) = 0\big)$ represents total ignorance as the evidence doesn't allow us to conclude whether A is true or false.

The $N - \Pi$ diagram summarises the knowledge about an event A based on the pair of measures $\big(N(A), \Pi(A)\big)$, as shown in Figure 1. Points are only allowed on the axes $N = 0$ (tentative acceptance of $\bar{A}$) and $\Pi = 1$ (tentative acceptance of A), and other areas correspond to inconsistent possibility distributions (that is functions $\pi(x)$ defined in a manner that does not respect the axioms 1 and 2 or their consequences). Three points are particularly of interest: the more $N(A) \to 1$, the more certain event A is; the more $\Pi(A) \to 0$, the more certain $\bar{A}$ is; and the closer to $(N = 0, \Pi = 1)$, the more uncertain we are. We call the latter the ignorance point.

Possibility and probability have often been characterised as complementary theories that address different issues, but Dubois and Prade (2012) suggest that possibility measures can be viewed as bounds on imprecise probability measures. There can be multiple definitions of consistency (Delgado and Moral, 1987), but we follow Dubois et al. (2004) who held that a probability measure P and possibility measure Π are consistent if the probability of all events A satisfies $P(A) \leq \Pi(A)$. The definition of necessity implies that the probability $P(A)$ is likewise bounded from below by the necessity measure:

$$N(A) \leq P(A) \leq \Pi(A). \qquad \text{(Equation 1)}$$

Finally, we say that a possibility distribution π is at least as specific as another π' when $\pi(s) \leq \pi'(s)\ \forall s \in \mathcal{X}$, in which case π' is more conservative (or less informative) than π. Generally speaking, possibility theory is driven by the principle of minimal specificity, which states that we cannot rule out an hypothesis not known to be impossible (Dubois and Prade, 2012).

2.2 From data to possibility distribution

Let us consider a stochastic variable $x \in \mathcal{X}$ for which we try to make a prediction. The available evidence about x is a set $S = \{x_1, \ldots, x_{N_s}\}$ of N_s samples of x. To turn this information into a possibility distribution describing the knowledge on the actual value of x, we use the technique described by Masson and Denœux (2006). Their methodology is specifically designed to derive a possibility distribution from scarce raw data, and assumes that the data in S have been randomly generated from an unknown probability distribution P. The idea is, after binning the x-axis into n bins, to recover the simultaneous confidence intervals at level $1-\beta$ 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. $\Pi(A) \geq P(A)\ \forall A$ in $100\beta\%$ of the cases). The simultaneous confidence intervals for multinomial proportions are computed by means of the formulation of Goodman (1965) (presented in Appendix B). Other formulations such as the imprecise Dirichlet model of Walley (1996) exist. However both models do not provide the same guarantees: Goodman's formulation provides multinomial confidence intervals at level β for the physical 'true' multinomial probabilities $\{p_i, i = 1, \ldots, n\}$—according to the classification of probabilities by Good (1966). The imprecise Dirichlet model, characterised by a parameter s, provides intuitive, logical probabilities (Walley, 1996) instead: namely, the upper and lower bounds on the probability of a given event A represent rational beliefs and rational betting rates that are justified by the evidence at hand. In this work, we only consider the Goodman's formulation. Appendix A presents Masson & Denoeux's technique step by step.

The above stage is essential for our application, especially in the case of a system with a limited sample set S. Indeed, the classical approach for the probability-possibility transformation proposed by Dubois et al. (1993) directly uses the vector of

frequencies $\{n_i/N_s, i = 1,\ldots,n\}$ as the true vector of probabilities $\{p_i, i = 1,\ldots,n\}$. The uncertainty on the p_i that is due to the limited size of S is therefore not taken into account. For our application, seeking guarantees on the possibility of observing an event of interest, it is necessary to account for such uncertainty.

One could observe that the above computations of possibility distributions mostly rely on probabilities. So why should we withdraw from the qualifying term 'probabilistic'? Since the principle according to which what is probable must first be possible was stated by Zadeh (1978), quantitative interpretations of possibility distributions have been connected to probability theory and transformations from one to the other have been developed. Thus, possibility distributions, as fuzzy membership functions, can be seen as encoding a family of nested confidence intervals (Dubois and Prade, 1982). More generally, De Cooman and Aeyels (1999) have shown that possibility measures encode families of probability distributions. As shown by Equation 1, a possibility distribution can be seen as a complete and consistent framework to deal with imprecise probabilities. It 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 incompleteness of data to be accounted for, while a point probability hides the fact that the said probability cannot be fully trusted. Although possibility distributions are connected to probabilities, they consequently provide a very different representation of the knowledge at hand, that belongs to the field of imprecise probabilities.

2.3 From possibility distribution to prediction

In this study, we focus on the binary interpretation of π, while the continuous interpretation is developed in Le Carrer (n.d.). We are consequently interested in the prediction of an event A of interest.

According to Section 2.1, we can extract from π the possibility $\Pi(A)$ and necessity $N(A)$. Such measures provides coordi-es to locate the corresponding point $\mathcal{P}$ in the $N-\Pi$ diagram sketched in Figure 1. Recall that the closer $\mathcal{P}$ is to the point $(1,1)$, the more necessary A becomes. The closer $\mathcal{P}$ is to the point $(0,0)$, the less possible it becomes. When $\mathcal{P}$ is around $(0,1)$, the user is in situation of ignorance: the information at hand does not justify a conclusion about A. One way of making predictions is consequently to use a threshold on either Π, N, or a function of both. However, using Π or N only would loose information. The credibility $C(A) = \frac{N+\Pi}{2}$ was introduced by Liu (2006) to address this issue. Thresholds $p_t \in [0,1]$ can thus be used to make predictions: $C(A) \geq p_t \Rightarrow A$ predicted. Similarly to the probabilistic approach, such thresholds can be selected by means of a Relative Operating Characteristic or a Precision-Recall Curve, in order to fit the constraints provided by the user (e.g. relative level of false alarms). More generally, any functional $P_\alpha = \alpha N + (1-\alpha)\Pi$, $\alpha \in [0,1]$ allows to reduce the interval on $P(A)$ (cf. Equation 1) into a point-prediction $P_\alpha(A)$. Although information is lost, this may be more convenient

for decision-making. α is then chosen so as to optimise a performance metric designed for probabilistic predictions, over a test set.

Finally, we propose another interpretation, following directly the axioms of possibility theory and their consequences (cf. Section 2.1). Since $N(A) > 0$ means tentative acceptance of A with confidence $N(A)$ (lower bound on $P(A)$, bounded on top by $\Pi(A)$), and conversely $\Pi(A) < 1$ means tentative acceptance of $\bar{A}$ with confidence $1 - \Pi(A)$, we can develop the following logic:

- $N(A) > 0$ implies A is predicted, with associated probability $N(A)$ (risk prone and risk neutral) or $\Pi(A)$ (risk averse) ;
- $\Pi(A) < 1$ implies $\bar{A}$ is predicted, with associated probability $N(\bar{A}) = 1 - \Pi(A)$ (risk averse and risk neutral) or $\Pi(\bar{A}) = 1 - N(A)$ (risk prone) ;
- $\big(N(A) = 0, \Pi(A) = 1\big)$ implies that either A (risk averse) or $\bar{A}$ (risk prone) is predicted with associated probability P_{IGN} (resp. $1 - P_{IGN}$). In practice, $P_{IGN} = 0.5$ (typically in the situation of no prior information) or P_{IGN} is defined with the observed frequency of A among points falling in the ignorance area.

In the so-called risk neutral case, the lower bound on $P(A)$ (resp. $P(\bar{A})$, that is the confidence level on observing A (resp. $\bar{A}$), is used as associated probability. More generally, the risk-prone and risk-averse predictions outside of ignorance can be encoded as such:

- $N(A) > 0$ implies A is predicted, with associated probability $P_\alpha(A)$;
- $\Pi(A) < 1$ implies $\bar{A}$ is predicted, with associated probability $P_\alpha(\bar{A}) = 1 - P_\alpha(A)$,

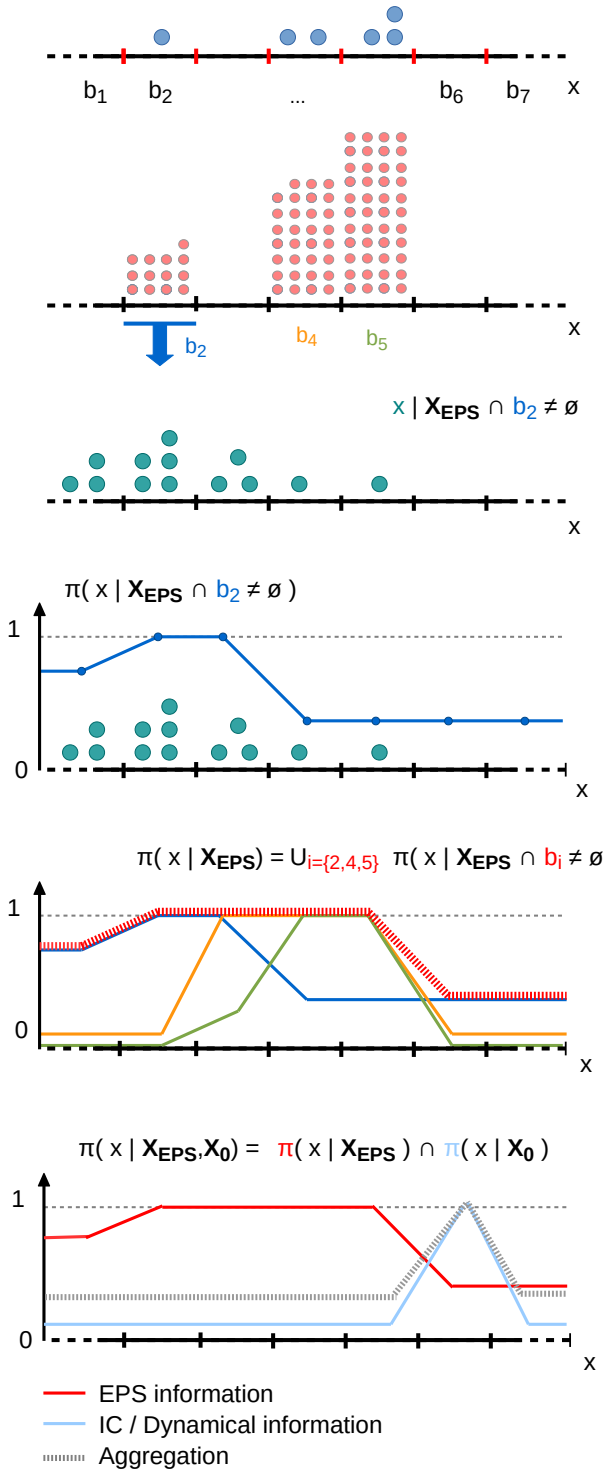
where $\alpha \to 0$ (risk averse), $\alpha \to 1$ (risk prone).

Thereafter, we name pred-CRED the credibility approach, pred-ALPHA-α the P_α approach (note that pred-CRED is in practice equals to pred-ALPHA-0.5) and pred-TENT-AV (resp. pred-TENT-PR and pred-TENT-NEU) for the risk-averse tentative approach (resp. risk-prone and risk-neutral tentative approaches).

3 Framework

3.1 Notations and information at hand

We are interested in the prediction of the state variable x_{t_0+t} of a dynamical system $\mathcal{S}$ at lead time t, starting from the IC x_{t_0}. $x \in \mathbb{R}$ refers to the component of interest of $\mathcal{S}$ (if directly accessible), or to a function of the inaccessible component of interest, measured in the model space. We call *verification* the actual value of x_{t_0+t}.



1 The axis is binned and EPS members are placed in the bins.

2 For each bin b_i occupied by at least one member of the EPS $\boldsymbol{X_{EPS}}$, we collect the EPS members of the archive l_t that fell in the same bin at that same lead time t.

3 For each occupied bin b_i, we collect the verifications associated with the above subset of archived EPS members and place them in the bins over the axis.

4 We compute from this set of N_S analogs the possibility distribution describing the system state x at lead time t, given that a member of $\boldsymbol{X_{EPS}}$ has fallen in bin b_i.

5 The possibility distribution for the system state at a given lead time, given the EPS, is the union (i.e. envelope) of the possibility distributions associated to each occupied bin.

6 To take into account the initial conditions $\boldsymbol{X_0}$ (IC) and local dynamics of the system, we intersect this possibility distribution with a possibility distribution based only on ICs, possibly expanded through delay embedding if we dispose of a long enough record of the system.

Figure 2. Step by step illustration of our framework.

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: $\tilde{\boldsymbol{x}}_{\boldsymbol{t_0+t}} = \{\tilde{x}^1_{t_0+t}, \ldots, \tilde{x}^M_{t_0+t}\}$.
2. An archive $\mathcal{I}_t$ containing the pairs $\left(\tilde{\boldsymbol{x}}_{t_k+t}, x_{t_k+t}\right)$ for the lead time t of interest and N_I different starting time t_k , $k = 1, \ldots, N_I$. These instances are chosen so that the initial points x_{t_k} and $x_{t_{k+1}}$ of two successive trajectories are statistically independent from each other (namely, in our model example, they are spaced of 3 time units, that is about 15 days, well above ≈ 1 day, the first minimum of the mutual information between x_t and $x_{t+\tau}$).
3. A time series of (preferably continuous) N_{I_A} past observations of x, that we denote $\mathcal{I}_A$, containing the IC x_{t_0} of interest.

3.2 Deriving possibility distributions from EPSs

The objective of our possibilistic interpretation of EPSs is to derive from an EPS $\tilde{\boldsymbol{x}}_{\boldsymbol{t_0+t}}$ and the archive $\mathcal{I}_t$ a possibility distribution $\pi(x_{t_0+t}|\tilde{\boldsymbol{x}}_{\boldsymbol{t_0+t}}, \mathcal{I}_t)$, that would encode the knowledge derived from the EPS about the verification x_{t_0+t} at a given lead time t. For readibility, we omit to indicate $\mathcal{I}_t$ in the upcoming equations, however the possibility distributions are derived from this source of information combined with the EPS at hand. The procedure described in this section is summarised and illustrated in the steps 1—5 of Figure 2.

Both system and model being (to a certain extent) deterministic and stationary or 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_0+t} associated with a small range S_x of the values taken by ensemble members, then a future verification x_{t_0+t} should belong to that set of possible values when an ensemble member $\tilde{x}^m_{t_0+t}$ falls within S_x. Beyond that, we would like to know which one of these values are more possible than others for x_{t_0+t}. In other words, we would like to estimate the possibility distribution $\pi(x_{t_0+t}|\tilde{x}^m_{t_0+t} \in S_x)$. 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 will not affect the resulting possibility distribution for x_{t_0+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 the $|B|$ possibility distributions $\pi(x_{t_0+t}|\tilde{x}^m_{t_0+t} \in b_j)$ where $b_j \in B$. Namely, following the methodology presented in Section 2.2, for each bin $b_j \in B$ occupied by at least one ensemble member $\tilde{x}^m_{t_0+t} \in \tilde{\boldsymbol{x}}_{\boldsymbol{t_0+t}}$, we retrieve all the ensemble members from the archive $\mathcal{I}_t$ with index k such that $\tilde{x}^m_{t_k+t} \in b_j$, and build an histogram of the set of corresponding verifications x_{t_k+t} (called *analogs*) over the same partitioning of the x-axis, $\{b_i, i = 1, \ldots, n\}$.

The procedure above computes $|B|$ possibility distributions $\pi(x_{t_0+t}|\tilde{x}^m_{t_0+t} \in b_j)$, each dominating with a confidence $1-\beta$ the true probability distribution $P(x_{t_0+t}|\tilde{x}^m_{t_0+t} \in b_j)$ (i.e. verifying Equation 1 with confidence β). Each possibility distribution provides the possibilities for the verification x_{t_0+t} given the presence of one or more ensemble members in bin b_j. Each one is thus a partial view on the state x_{t_0+t}. Since there is only one truth for x_{t_0+t} (the system's actual state), we can merge them through a union operator (OR). Fuzzy set theory offers several definitions for computing the distribution resulting of the union of two fuzzy distributions. We adopt here the standard definition for its intuitive rationale: $\pi_{A\cup B}(x) = \max\big(\pi_A(x), \pi_B(x)\big)$.

We construct the resulting possibility distribution as:

$$\begin{aligned}\pi_{EPS}(x_{t_0+t} \in b_i|\tilde{\boldsymbol{x}}_{\boldsymbol{t_0+t}}) &= \bigcup_{j|b_j\in B} \pi(x_{t_0+t} \in b_i|\tilde{x}^m_{t_0+t} \in b_j) \\ &= \sup_{j|b_j\in B} \pi(x_{t_0+t} \in b_i|\tilde{x}^m_{t_0+t} \in b_j),\ i=1,\ldots,n. \end{aligned} \quad \text{(Equation 2)}$$

Observe that at this stage, we have not yet taken the ICs x_{t_0} into consideration in the selection of the analogs. In other words, π_{EPS} is too conservative due to a lack of information about the dynamics of $\mathcal{S}$ at the time of interest. To alleviate this issue, we consequently combine our framework to the empirical dynamic modelling of $\mathcal{S}$, that is to the reconstruction of its shadow attractor. More generally, any method providing dynamical analogs can be used.

3.3 Taking dynamical information into account

3.3.1 Attractor reconstruction

The procedure of attractor reconstruction consists for a dynamical system characterised by a variable x_t in finding the time delay τ and embedding dimension m such that the time delay vectors $\boldsymbol{x}_t = \big(x_t, x_{t-\tau}, \ldots, x_{t-(m-1)\tau}\big)$ allow to reconstruct the fully unfolded shadow attractor $\mathcal{A}'$ in the embedding space (that is such that no two distinct trajectories cross). We use the simplex projection method (Sugihara and May, 1990; Deyle and Sugihara, 2011; Sugihara et al., 2012), specifically designed when the attractor is used for prediction purposes. The idea is to find the couple (m,τ) that maximises the correlation between verification and prediction, where the prediction of the future state of the system is given by a weighted mean of n_A analog trajectories. In other words, given the IC of interest $\boldsymbol{x}_{t_0}$ in the phase space, we find the n_A closest neighbors (in the sense of the Euclidean L2 norm), and follow their trajectories up to lead time t. This provides us with the desired n_A analogs.

Again, any similarity-based method providing dynamical analogs (that is taking into account information on the ICs, where IC is understood as the point IC x_{t_0} or as a longer vector containing dynamical information) can be used to provided the n_A analogs.

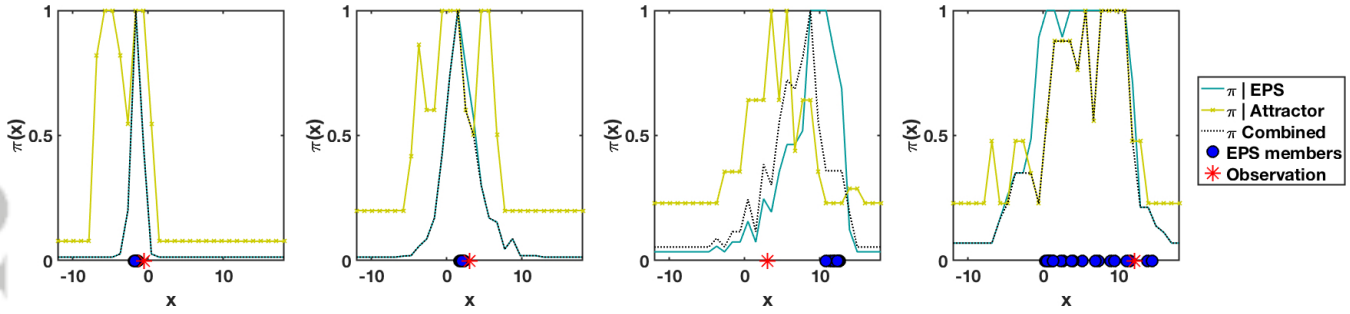


Figure 3. EPS- and attractor-based possibility distributions and their combination at lead times $t = \{1, 3, 5, 7\}$ days (left to right).

3.3.2 Converting dynamical analogs into a predictive possibility distribution

Depending on the archive $\mathcal{I}_A$ at hand and the embedding dimension selected for the reconstruction, the attractor can be more or less dense, especially in the areas of rare events. We consequently avoid analog-based point predictions, and again resort to possibility distributions to extract the information given by the analogs. This allows us to account for sparse analog datasets and ensure that non-homogeneous density in the phase space does not blur results. Thus, we follow the procedure described in Section 2.2 to draw the possibility distribution $\pi_{DYN}(x_{t_0+t}) = \pi(x_{t_0+t}|\boldsymbol{x}_{t_0}, \mathcal{I}_A)$ for the verification x_{t_0+t} associated with the IC $\boldsymbol{x}_{t_0}$ in the phase space.

3.3.3 Combining EPS and dynamical information

π_{EPS} and π_{DYN} are two views on the actual system state x_{t_0+t} that are both supposed to be complete, although possibly too conservative, due to their limited and imperfect source of information about the state of the system. We consequently combine them in an AND manner: $\pi(x_{t_0+t}|\tilde{\boldsymbol{x}}_{\boldsymbol{t_0+t}}, \boldsymbol{x}_{t_0}) = \pi_{EPS} \cap \pi_{DYN}$, which we posit should alleviate their respective over-conservatism. The intersection of two possibility distributions is classically given by Zadeh's extension principle (Zadeh, 1978; Hose and Hanss, 2019):

$$\pi_{A\cap B}(x) = \inf\left(\pi_A(x), \pi_B(x)\right). \qquad \text{(Equation 3)}$$

The final (a.k.a. combined) possibility distribution is consequently:

$$\pi_{COMB}(x_{t_0+t} \in b_i|\tilde{\boldsymbol{x}}_{\boldsymbol{t_0+t}}, \boldsymbol{x}_{t_0}) = \inf\left(\pi_{EPS}(x_{t_0+t} \in b_i), \pi_{DYN}(x_{t_0+t} \in b_i)\right),\ i = 1, \ldots, n. \qquad \text{(Equation 4)}$$

The resulting distribution is finally normalised to one, to verify axiom 1 from Section 2.1. This consists in using the following transformation, for a generic possibility distribution $\pi(x)$:

$$\pi(x) \leftarrow \begin{cases} \frac{\pi(x)}{\max_x(\pi(x))} & \text{if } \max_x(\pi(x)) > 0 \\ 1 & \forall x, \text{ otherwise} \end{cases}. \quad \text{(Equation 5)}$$

In practice, if the min-envelope defined by Equation 4 is null everywhere (typically when both EPS- and IC-based distributions are peaked with non-overlapping support), we turn it into a uniform distribution. The philosophy behind is that independent sources of information are contradictory so we are in a situation of ignorance (everything is possible). Otherwise, we divide the min-envelope by its maximum, to get a distribution satisfying the axioms of possibility theory (see axioms 1 and 2, namely: something must be possible within the universe of the variable of interest). The philosophy behind is that the maximum of the min-envelop corresponds to area(s) with the highest joint support of EPS- and IC-based sources of information. Since at least something must be possible (cf. above-mentioned axioms of definition), these areas are associated to a possibility measure of 1 and other events scaled accordingly. An illustrative example is provided Figure 3.

3.3.4 Guarantees

We conclude this section with a focus on the formal guarantees that our methodology provides. By construction, the possibility distributions π_{EPS} and π_{DYN} dominate with a given confidence level β (in the case of Goodman's formulation) the true probability distribution of the future x_t. Their joint aggregation is designed to make the resulting possibility distribution more specific. Although such a step cannot in general maintain the same level of confidence regarding the property $P(A) \leq \Pi(A)\ \forall A$ [1], π_{COMB} still provides guarantees when it comes to the lower bound of $\Pi(A)$. Indeed, from axiom a. of Section 2.1, if $x_t = x^*$ is actually observed, we have: $\pi_{EPS}(x^*) > 0$ and $\pi_{DYN}(x^*) > 0$. Consequently, by definition of the combined possibility distribution (Equation 4), $\pi_{COMB}(x^*) > 0$ as well. Thus, the guarantee $\Pi(A) > 0$ when $x^* \in A$ is maintained. This allows risk-averse decision-makers to get a guarantee about the possibility of observing A: all observations of A are associated to a non-null $\Pi(A)$. However, taking precautionary action whenever $\Pi(A) > 0$ is not always feasible for economical reasons. In such a case, the AND-fusion of π_{EPS} and π_{DYN} allows to reduce the basis level γ such as $\pi_{COMB}(x) \geq \gamma$, $\forall x$, and consequently to increase the upper bound on the necessity, $N(A) \leq 1 - \gamma$, $\forall A$, that is the minimal confidence level in favor of A.

[1] Hose and Hanss (2019) discusses this point and shows how using the so-called general aggregation ensures that the consistency between probability and possibility measures is maintained, whatever the level of interaction, or dependence, between the variables at hand.

The decision maker can then use it to judge whether the possible event A is actually more or less probable. The evaluation of the formal guarantees associated to our framework is developed in Le Carrer (n.d.).

4 Experimental setting

4.1 Test bed: the imperfect L96 system

We reproduce the experiment designed by Williams et al. (2014), who used an imperfect L96 model to investigate the performances of ensemble postprocessing methods for the prediction of extreme events. The system dynamics is governed by the following system of coupled equations, where the X variables represent slow-moving, large-scale processes, while Y variables represent small-scale, possibly unresolved, physical processes:

$$\frac{dX_j}{dt} = X_{j-1}(X_{j+1} - X_{j-2}) - X_j + F - \frac{hc}{b}\sum_{k=1}^{K} Y_{j,k} \quad \text{(Equation 6)}$$

$$\frac{dY_{j,k}}{dt} = cbY_{j,k+1}(Y_{j,k-1} - Y_{j,k+2}) - cY_{j,k} + \frac{hc}{b}X_j \quad \text{(Equation 7)}$$

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. An imperfect version of the L96 system is implemented to generate predictions for the variables X_j. In Equation 6, $-\frac{hc}{b}\sum_{k=1}^{K} Y_{j,k}$ is replaced with a quartic polynomial in X_j:

$$\dots 32 - 1.262X_j + 0.004608X_j^2 + 0.007496X_j^3 - 0.0003226X_j^4 \quad \text{(Equation 8)}$$

To reproduce the perturbation of the ICs, each perturbed variable $\tilde{X}_j$ is randomly and independently drawn from $\mathcal{N}(X_j, 0.1^2)$. M members are thus sampled independently around the true value of X_j. The ensemble predictions are initialised each time a new trajectory record starts, and integrated forward in time up to the 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). A lead time of 0.2 model time units after initialisation is noted $t = 1$ and can be associated with approximately 1 day in the real world (Lorenz, 1996).

In the following, we adopt a monovariate perspective, that is we consider each dimension of the model space independently. More specifically, we illustrate our methodology with predictions of the variable X_1.

4.2 Reference models: Gaussian ensemble dressing and raw EPS distribution

In many cases, the statistical postprocessing of EPSs generates forecasts in the form of predictive probability distributions $p(x_{t_0+t}|\tilde{\boldsymbol{x}}_{\boldsymbol{t_0+t}},\theta)$, where $\tilde{\boldsymbol{x}}_{\boldsymbol{t_0+t}} = \{\tilde{x}^1_{t_0+t},\ldots,\tilde{x}^m_{t_0+t}\}$ is the ensemble, θ a vector of parameters and p a (sum of) parametric distribution(s). Bayesian model averaging distributions (BMA; Raftery et al. (2005)) are weighted sums of M parametric probability distributions, each one centered around a linearly corrected ensemble member. In this work, the members are exchangeable, so the mixture coefficients and parametric distributions do not vary between members and the BMA boils down to an ensemble dressing procedure. We compare our method (referred to as EPS, DYN$-m$ or COMB$-m$ whether we use π_{EPS}, π_{DYN} or π_{COMB}, with $-m$ specifying the number of dimensions taken into account for the IC) against a Gaussian ensemble dressing, whose predictive probability distribution reads (Roulston and Smith, 2003):

$$p(x_{t_0+t}|\tilde{\boldsymbol{x}}_{\boldsymbol{t_0+t}})_\theta = \frac{1}{M}\sum_{i=1}^{M}\mathcal{N}(a\tilde{x}^i_{t_0+t}+\omega,\sigma^2) \qquad \text{(Equation 9)}$$

where $\mathcal{N}(\mu,v)$ is the normal distribution of mean μ and variance v. We infer the parameters $\theta = \{a,\omega,\sigma\}$ through the optimisation of a performance metric, here the ignorance score (Roulston and Smith, 2002), or negative log-likelihood, a strictly proper and local logarithmic score. To that end, we use the nonlinear programming solver provided by the software MATLAB® and apply the guidance developed in Bröcker and Smith (2008) to initialise the optimisation algorithm and provide robust solutions. Our training set contains N_I pairs {EPS,verification} for each lead time of interest $t = \{1,3,5,7\}$ days, that is the same information as the archive $\mathcal{I}$ used in our framework. To account for the variability of results from one testing set to the other, in the same line as Williams et al. (2014), we repeat the optimisation procedure 20 times on different samples. We then use the resulting 20 sets of parameters to compute the performance metrics relative to the probabilistic approach. Finally, we take the average of these 20 scores, that we report on the graphs as representative of the performances of the probabilistic approach.

In addition to the performances of the Gaussian ensemble dressing (hereafter GEB), we report the performance of probability distribution directly derived from the raw EPS (namely, an histogram normalised into a probability distribution). We refer to it as the RAW method.

4.3 Evaluation of performances

In this work, we have developed the binary interpretation of a predictive possibility distribution $\pi(x)$. Further work on the continuous interpretation and guarantees is presented in (Le Carrer, n.d.) by the authors. We consequently assess the predictive performance of our framework in the case of an extreme event: $A = \{x \leq q_5\}$, where q_5 is the quantile of order 5% of the climatic distribution of x. Such a choice allows us to target the issues of probabilistic interpretation of EPSs raised in introduction. To that end, we use two indicators commonly chosen for evaluating binary probabilistic predictions: the ignorance score and the precision-recall curves. We finally discuss reliability by means of reliability diagrams. These modalities of evaluation are presented below, along with the concept of U-uncertainty.

4.3.1 U-uncertainty

The U-uncertainty, also known as the generalized Hartley measure for graded possibilities (Klir, 2006), allows to measure the nonspecificity of the possibility distribution $\pi(x)$ at hand. In a continuous setting, it reads:

$$U(\pi) = \int_0^1 \log_2 |C_\pi^\alpha| d\alpha \qquad \text{(Equation 10)}$$

where $|C_\pi^\alpha|$ is the L_1 norm of the α-cut $C_\pi^\alpha = \{x \in \mathcal{X} | \pi(x) \geq \alpha\}$. Another way to compute it in a discretised setting is to order the possibility profile π in such a way that $1 = \pi_1 \geq \pi_2 \geq \ldots \geq \pi_n$ with $\pi_{n+1} = 0$ by definition. The following relationship then applies (Klir, 2006):

$$U(\pi) = \sum_{i=2}^{n} \pi_i \log_2 \frac{i}{i-1} \qquad \text{(Equation 11)}$$

$0 \leq U(\pi) \leq |\log_2 \mathcal{X}|$ defines the upper and lower bounds for a profile π over domain $\mathcal{X}$, obtained respectively for a Dirac-like profile and a uniform profile. Given two possibility profiles π and π', $U(\pi) \leq U(\pi')$ is equivalent to say that π is more specific (i.e. more informative) than π'.

This is not an indicator of prediction performance *per se*, however we will use it to discuss the information content of π_{EPS}, π_{DYN} and π_{COMB}.

4.3.2 Ignorance score

The ignorance score is designed to measure the skill of probabilistic predictions. It can be interpreted from an information-theory point of view in terms of the difference in expected returns that one would get by placing bets proportional to their probabilistic forecasts compared to bets that someone with perfect knowledge of the future would place. The empirical assessment of the ignorance score is the average over a test set of size N of the ignorance of each probabilistic prediction:

$$S_N(G) = \frac{1}{N}\sum_{i=1}^{N} -\log_2 G(O_i) \qquad \text{(Equation 12)}$$

where O_i is the event actually observed for sample i and $G(O_i)$ its predictive probability. In the probabilistic framework, S_N takes positive values only and each unit indicates an additional bit of ignorance on the forecaster's side.

The possibilistic framework does not provide a single probability $G(O_i)$ but a couple $\big(N(O_i),\Pi(O_i)\big)$ such that $N(O_i) \leq P(0_i) \leq \Pi(O_i)$ where $P(O_i)$ is the actual probability of event O for sample i. As described in Section 2.1, $N(A) > 0$ implies $\Pi(A) = 1$ and similarly $N(A) = 0$ (that is $\Pi(\bar{A}) = 1$) implies $\Pi(A) \leq 1$. In other words, whatever the verification O, a good possibility distribution π must derive into:

(A) $\Pi(O) = 1$

(B) $N(O) \geq 0$, with $N(O) \to 1$ preferred since it means that O is all the more necessary which makes the prediction less uncertain

An interesting way to extend the ignorance score to our possibilistic framework is to extract the credibility of the actual outcome from the couple possibility/necessity and use it as probability:

$$S_{N_\pi}(\pi) = \frac{1}{N}\sum_{i=1}^{N} -\log_2\left(\frac{N(O_i)+\Pi(O_i)}{2}\right) \qquad \text{(Equation 13)}$$

The score takes only positive values. Condition (A) is satisfied in average when $S_{N_\pi} \leq 1$ with condition (B) satisfied when $S_{N_\pi} \to 0$.

Both $N(O)$ and $\Pi(O)$ can be interpreted as predictive probabilities of the event O. One is (generally) an under-estimation and the second (generally) an over-estimation. The quantity $\frac{N(O_i)+\Pi(O_i)}{2}$ is consequently homogeneous to a probability and the score S_{N_π} has the same interpretation in terms of information theory as the classical ignorance score applied to the predictive probability $\frac{N+\Pi}{2}$. The choice of such a functional can be discussed, as there exist many other possible transformations to

reduce the couple $\left(N, \Pi\right)$ to a probability G. Beyond the classical $G(O) = \alpha N(O) + (1-\alpha)\Pi(O) = P_\alpha(O)$, where α can be optimised based on a performance metric, we do not discuss it in this work. We solely use this transformation with $\alpha = 0.5$ in order to get an ignorance score allowing to check easily whether properties (A) and (B) are verified in average, in addition to assess the information content of the derived predictive probability.

4.3.3 Precision recall curves

Traditionally, relative operating characteristics (ROCs) are used to estimate the ability of a predictive model to discriminate between event and non-event. Given a binary prediction (yes/no w.r.t. event A), the ROC plots the hit rate (fraction of correctly predicted A over all A observed) versus the false alarm rate (fraction of wrongly predicted A over all $\bar{A}$ observed).

However, when the dataset used to plot such characteristic is significantly imbalanced (the frequency of verification of A is significantly smaller than the frequency of verification of $\bar{A}$), the false alarm rate is biased towards lower values. Recent works, e.g. Saito and Rehmsmeier (2015), suggest to use instead precision-recall curves (PRCs). The precision (rate of correctly predicted A over all A predicted) is plotted as a function of the hit rate (a.k.a. recall, the terminology used in the machine learning research community). In other words, the false alarm rate is replaced with the precision. This removes any reference to the class that is not of interest ($\bar{A}$), which, when being the majority in an imbalanced dataset, biases the false alarm rate and consequently the conclusions that one could draw about prediction performances. Conversely, PRCs provide a more reliable prediction of the future classifier's performances. Our focus being on rare events, in this study characterised by a climatological frequency $c(A) = 0.05$, we consequently use PRCs to assess the predictive skills of our framework.

In both probabilistic and possibilistic cases, we use increasing thresholds $p_t \in [0, 1]$ for making the decision (A predicted if $P(A) \geq p_t$ (resp. $C(A) \geq p_t$) in the probabilistic (resp. possibilistic) framework) and report the associated precision and recall in the graph, forming a PRC. This allows us to compare the discrimination skill of both approaches.

4.3.4 Reliability diagram

This presentation of reliability diagrams draws on our previous work (Le Carrer and Green, 2020), where we first introduced our fuzzy and 3-dimensional versions of the metric. Reliability diagrams plot the observed conditional frequencies against the corresponding forecast probabilities for a given lead time. They illustrate how well the predicted probabilities of an event correspond to its observed conditional frequencies. The predictive model is all the more reliable (i.e. actionable) when the associated curve is close to the diagonal, which represents perfect reliability. The distance to the diagonal indicates underforecasting (curves above) or overforecasting (curves below). Distance above the horizontal climatology line (frequency of A over

the whole archive $\mathcal{I}$) indicates the resolution of the system, i.e. how well it discriminates between events and non-events. The cones defined by the no-skill line (half-way between the climatology and perfect reliability) and the vertical climatology line allow us to define areas where the forecast system is skilled.

This metric is obviously designed for probabilistic predictions. However, the possibility-probability equivalence (Equation 1) allows us to use it as well for possibilistic outputs and compare their actionability with purely probabilistic prediction schemes. To draw a standard reliability diagram from possibilistic predictions, we use the functional $P_\alpha(A)$, where α is discretized on $[0,1]$. For a given set of N_s predictions $(N(A),\Pi(A))$, for each $\alpha_i \in [0,1]$, the N_s $P_{\alpha_i}(A)$ are computed and a traditional reliability plot is drawn. Each α_i-plot indicates how using $P_{\alpha_i}(A)$ as a probability for A is reliable and actionable on the long term. Seen as a whole, this bounded set of reliability plots allows to characterise the reliability of the probabilities given through $N(A) \leq P(A) \leq \Pi(A)$.

.

5 Results & Discussion

We now characterise the predictive performances of our possibilistic framework and discuss them in comparison with the skill of the probabilistic reference approach. If not mentioned otherwise, all results presented in this Section use $n = 30$ bins to partition the $x-$axis [2], an archive of EPS/verification containing $N_I = 1560$ independent trajectories of length $t = 7$ days, and a continuous time series of x of length $N_{I_A} = 2.10^6$ sampled at the same frequency as the EPS trajectories. These are operational figures: an EPS-archive of such size N_I corresponds to 30 years of data, which corresponds to the standard length of a historical re-forecast dataset (Hamill et al., 2004; Hagedorn et al., 2008). The time series of length N_{I_A} above-mentioned roughly equals 55 years of system record, which for geophysical variables is reasonable. We will conclude by discussing the effect of N_{I_A} on performances. The calibration set (for parameter n_A) and test sets each consist in $N = 40.10^3$ independent trajectories of length $t = 7$ days and the corresponding EPS predictions. All EPSs have beforehand been preprocessed to remove the constant bias.

Finally, when it comes to the parameter β of the Goodman formulation, Masson and Denœux (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 $P\big(P(A) \leq \Pi(A)\big) \to 1 \ \forall A$. We consequently use

[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.

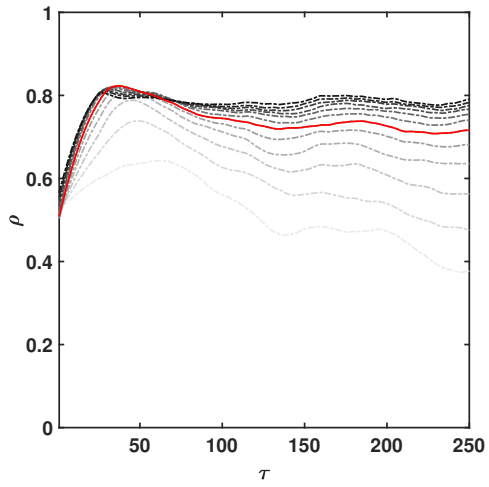


Figure 4. Results of the simplex method applied to the L96 system. The Pearson correlation coefficient between verification at lead time $t = 1$ day and the prediction computed by means of a weighted mean of the $m + 1$ closest analogues in the reconstructed phase space of embedding dimension m and time-delay τ. Each dashed curve corresponds to a different m, varying on $[4, 15]$. Larger m are darker. We top the plots with the solid red curve corresponding to the optimal or close to optimal m overall τ: $m = 9$.

$\beta = 0.9$ which, without impairing guarantees, tends to provide less conservative distributions as shown for the same case study in Le Carrer (n.d.).

5.1 Attractor reconstruction

The simplex method introduced in Section 3.3 is applied to the lead time $t = 1$ day from the continuous archive $x_{t_1}, \ldots, x_{t_{N_{I_A}}}$ of length $N_{I_A} = 2.10^6$ and time step similar to the EPS's time resolution. A clear optimum is found for the couple $(m = 9, \tau = 37)$ (cf. Figure 4). Herafter, when m is not explicitly mentioned for methodologies COMB-m or DYN-m, the reader will understand that $m = 9$.

5.2 Setting the number of analogs n_A

As illustrated in Figure 5, the parameter n_A plays an important part in the shape of π_{DYN} and a careful calibration is consequently recommended. Figure 5 shows the effect of increasing the number of analogs $n_A \in \{10, 50, 100, 500, 1000\}$ on π_{DYN}. We observe that increasing n_A produces a more and more specific distribution by increasing the minimum confidence level $N(A) = 1 - \max_{x \notin A} \pi(x)$ about an event A in the peak area. Globally, $n_A = 100$ already provides interesting predictive information, however $n_A = 500$ may provide a better decision tool due to higher confidence levels in the peaks. We can wonder whether this higher confidence, artificially induced by a larger analog set, will prevent the detection of small tenden-

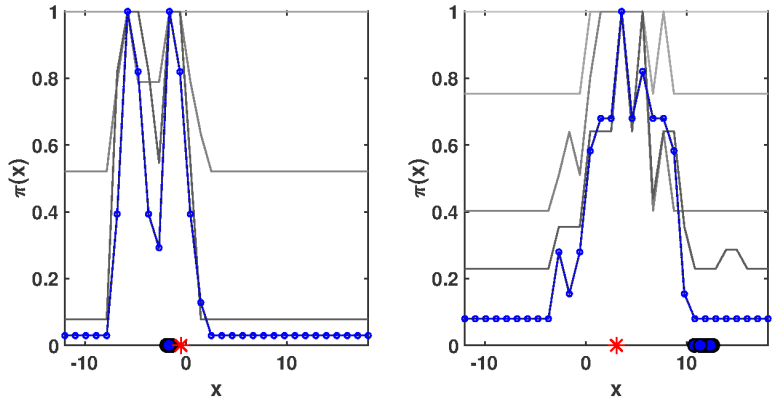


Figure 5. Effect of varying the number of analogs $n_A \in \{10, 50, 100, 500, 1000\}$ (darker lines for larger n_A) for lead times $t \in \{1, 5\}$ days (from left to right) on π_{DYN}. The distribution is highlighted with dots and color for $n_A = 1000$. The smaller n_A, the more conservative the distribution. Associated EPS members are marked as blue dots and the verification as a red star.

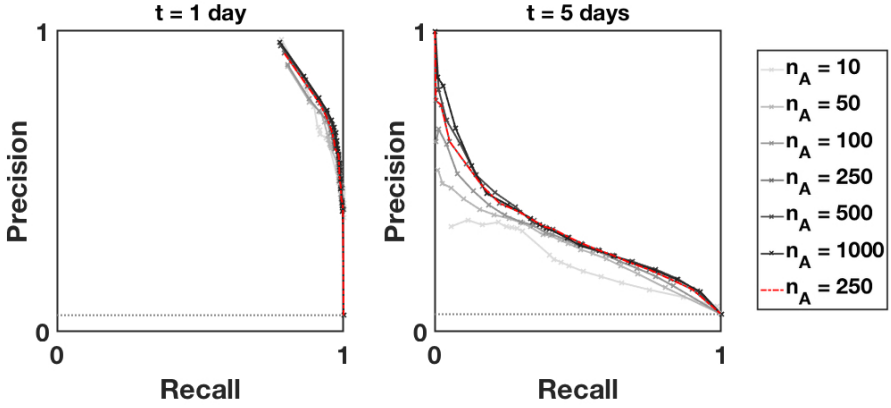


Figure 6. Effect of varying the number of analogs $n_A = \{10, 50, 100, 250, 500, 1000\}$ on the precision-recall curves at lead times 1 and 5 days. The darker the line, the higher n_A.

cies (typical of rare events). In particular, we consider the n_A closest neighbours around the IC $\boldsymbol{x}_{t_0}$, which does not imply that they are actually close, if the attractor is not dense in the area of interest. Figure 6 shows the effect of varying n_A over $\{10, 50, 100, 250, 500, 1000\}$ on the PRC, for lead times $t = \{1, 5\}$ days.

We observe that the performances in terms of PRC improve with growing n_A, yet they quickly converge to a maximum ($n_A \geq 250$). The sensitivity to n_A is more pronounced when the lead time increases. Such a convergence means that even though we integrate more distant analogs, the possibilistic methodology does not use this additional information in terms of density (which would dilute the information given by the closest analogs). Instead, the possibilistic interpretation of the

analog set is preserved. Globally $n_A = 250$ allows to get the best performances over the whole range of recalls, confirming the preliminary observations in Figure 5. We continue our experiments with this value for n_A.

5.3 Predictive performances

5.3.1 Information content

Figure 7 represents the empirical ignorance score for lead times varying from 1 to 7 days of the methods GEB, RAW, EPS, DYN-9, COMB-1 and COMB-9, broken down between its extreme event (EE) and non-extreme event (NEE) components, that is the average empirical ignorance for observed EE (resp. NEE) only. Note that due to the very small proportion of EE compared to NEE, the global empirical ignorance score is similar to the NEE's. For explanatory purposes, we represent as well the effect on the COMB possibility distributions of the aggregation method. Namely, we compare COMB-Z, using Zadeh's aggregation, defined in Section 3.3.3, to COMB-A, using the general aggregation defined in Hose and Hanss (2019) [3] and supposed to ensure the validity of the consistency principle (Equation 1) whether there is stochastic dependence or not between the variables to be fused. Finally, we compare the results for the possibility-based probabilities derived from the methodology pred-CRED, pred-TENT-NEU, and pred-TENT-AV and pred-TENT-PR with varying α. Note that the extreme versions of the last two ($\alpha = 0$ and $\alpha = 1$ respectively) cannot be directly used with the absolute ignorance score as for the risk-averse approach (resp. risk-prone) the NEE (resp. EE) component gets an infinite score. Indeed, if we take the risk-averse case (resp. risk-prone case), null probabilities are attributed to $\bar{A}$ (resp. A) whenever $\Pi(A) = 1$ (resp. $N(A) = 0$), which leads to infinite negative log-likelihood items. Conversely, using $0 < \alpha < 1$ ensures finite log-likelihood scores.

We first describe the results for possibility-based probabilities derived by means of the pred-CRED methodology. The NEE ignorance is slightly lower for probabilistic methods (GEB, RAW) than it is for the possibilistic approaches (EPS, COMB-1, COMB-9). However, when it comes to the case of interest, namely EE, the ignorance is significantly lower for the possibilistic approaches than for the probabilistic ones (where GEB shows that postprocessing improves the RAW result). The differences grows with the lead time.

If we analyse more in detail the possibilistic approaches, we note that in the NEE case, for lead times above 3 days, the aggregation of information (EPS and DYN) allows to lower the level of ignorance, all the more than the information about dynamics is refined (i.e. that the number of dimensions m taken into account to characterise the ICs is high). However, in the

[3] For N marginal possibility distributions $\pi_{X^k}, k = 1, \ldots, N$ about the variable $x \in \mathcal{X}$, the joint possibility distribution is defined as:

$$\pi_{X^1,\ldots,X^N}(x) = \min_{k=1,\ldots,N} \min\left(1, \pi_{X^k}(x)\right) \quad \forall x \in \mathcal{X}. \qquad \text{(Equation 14)}$$

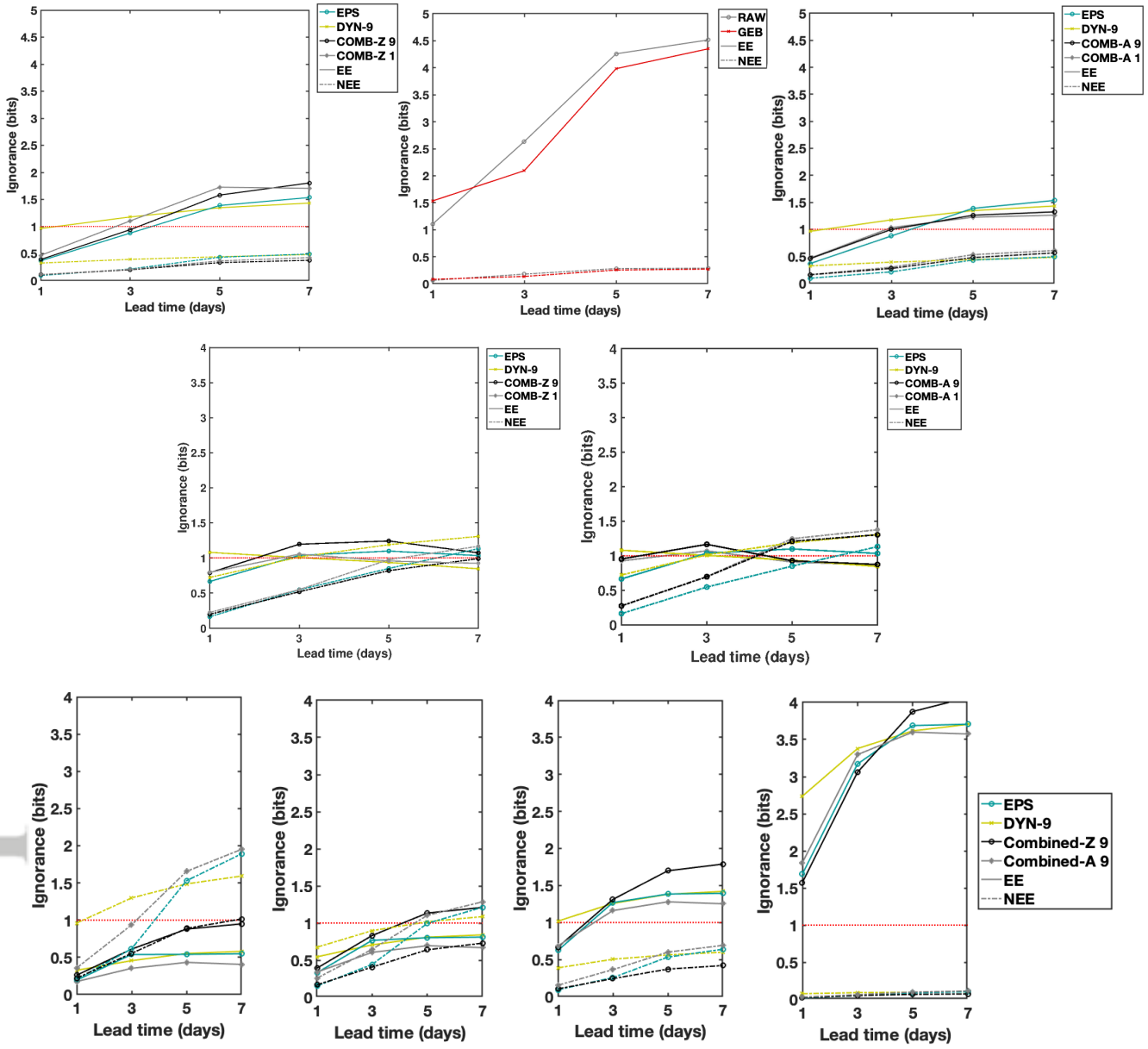


Figure 7. Empirical ignorance score of the methods described in the text. The upper plots use the pred-CRED approach to derive probabilities from possibility distributions. The middle plots use the pred-TENT-NEU and the lower plots use the pred-TENT-α with $\alpha \in \{0.1, 0.25, 0.5, 0.9\}$ from left to right. A dotted horizontal red line is plotted at 1 bit to visualise how guarantees are verified by possibilistic methodologies. In both first cases (top and middle), the left-most panels use the COMB-Z aggregation method while the right-most panels use the COMB-A approach for aggregation.

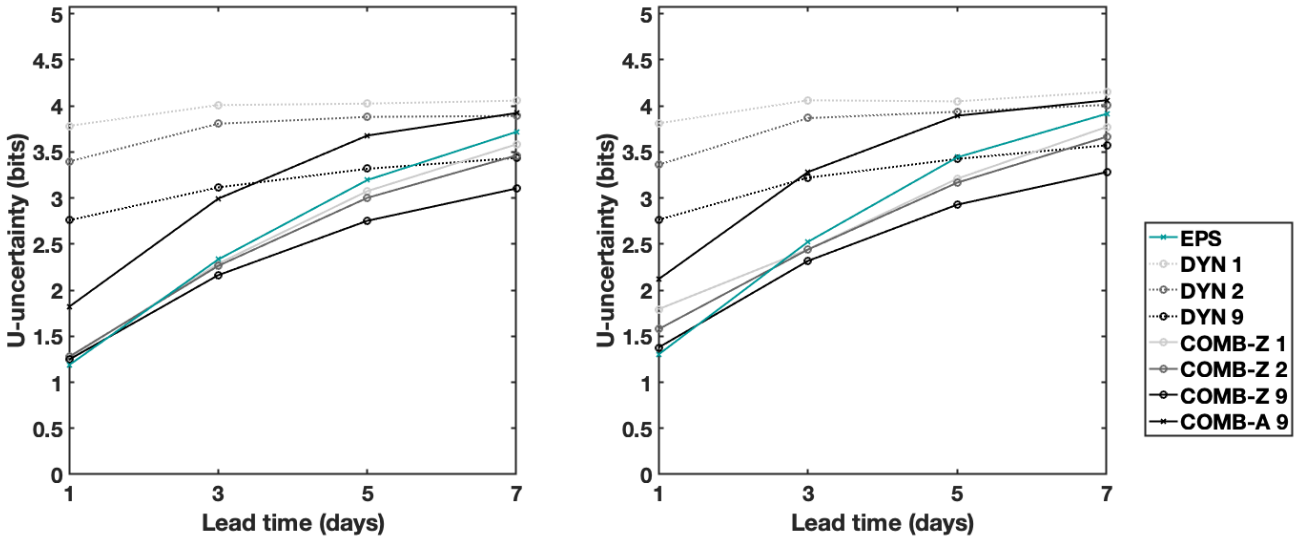


Figure 8. Average U-uncertainty of the possibility distributions described in the text for both NEE (left) and EE (right). The upper bound given the domain of definition of the variable at hand is $\log_2 |\mathcal{X}| = 5.08$, which would be obtained for a uniform possibility distribution.

EE case, the aggregation of information slightly increases the ignorance, even for small lead times. This is all the more true than the dynamical information is partial (i.e. m low), at least for lead times below 7 days.

Figure 8 allows to shed some light on this counter-intuitive observation. It shows that fusing the dynamical and EPS-based possibility distributions provides distributions that are more specific than both initial distributions at lead times above 1 day. Whether for NEE or for EE only, the effect is all the more marked than the lead time increases and the dynamical (and consequently the combined) possibility distributions are all the more specific than the information characterizing the ICs is complete (large m). If COMB distributions are more specific than EPS's and yet their information content is lower (their ignorance score is higher), it means that, in plain words, 'they missed their target' and led to situations such as $\big(N(A) = 0, \Pi(A) < 1\big)$ which means tentative acceptance of the complementary event $\bar{A}$ at level $1 - \Pi(A)$. And indeed, we note that the condition (A) is not verified in average for lead times above 3 days, since the empirical ignorance overpass 1 bit.

Using a different kind of aggregation, namely the general aggregation, allows to have COMB distributions more informative than the EPS ones in the case of EE, but not in the NEE case. This type of aggregation is indeed much more conservative as shown on Figure 8, which for EE is interesting but is less for more common events.

The pred-TENT-NEU methodology leads to EE results improved at larger lead times (below or closer to the 1-bit guarantee), especially in the EPS and COMB cases. However, results are significantly deteriorated for NEE, especially at large lead times.

It shows the potential of the methodology for risk-averse users, as conditions (A) and (B) are almost perfectly satisfied for both EE and NEE.

Finally, the last row of Figure 7 shows the effect of varying α in the pred-TENT-AV / pred-TENT-PR methodology. A small $\alpha \approx 0.1$ guarantees that the conditions (A) and (B) are met for EE, with best results for the distribution COMB-A. However only methodology COMB-Z, less conservative, allows to verify conditions (A) and (B) for both EE and NEE. As could be expected, increasing α leads to predictions less risk-averse, which increase the performances for NEE yet at the expense of EE's. One can however note that it exists a trade-off α where the ignorance score of such possibility-based predictions remains equal or better to the ignorance score of the probability-based predictions for NEE and EE simultaneously.

5.3.2 Ability to discriminate

Figure 9 gathers the PRCs of both predictive frameworks for lead times $\{1,3,5,7\}$ days. To gain insight, we report the PRCs obtained from the EPS, DYN and COMB-Z-9 possibility distributions. The PRCs are computed for $P_{\alpha=0} = \Pi$, $P_{\alpha=1} = N$ and $P_{\alpha=0.5} = 0.5(N + \Pi)$. We observe that using N as decision tool allows only small hit rates, especially when the lead time grows. Conversely, using Π doesn't allow small hit rates. Intermediate pooling such as $P_{\alpha=0.5}$ allows to cover the whole range of hit rates. Overall, π_{EPS} performs similarly to the probabilistic frameworks (points overlay) for $t \geq 3$ days, and even significantly better in the case of small recalls for $t = 7$ days. For smaller lead times, it performs slightly less well than the probabilistic approaches. In all three cases, π_{DYN} is significantly less successful than the latter for small and medium lead times. It becomes as interesting as them only from $t = 5$ days. The combined possibility distribution is consequently slightly below the probabilistic approach in terms of discrimination ability for small lead times, and becomes more interesting than the latter for $t \geq 5$ days.

We note than the performance of π_{COMB} is different than the performance of its best component (either π_{EPS} or π_{DYN}). At small lead times, it remains close to π_{EPS} performance, while at larger lead times, it goes beyond both. Combining both distributions in an AND manner consequently provides more predictive information than any single one of them contains.

These results can be explained by means of Figure 3. For short lead times, π_{EPS} is generally quite narrow (model error is low) and peaks around the true verification. Using it for prediction leads to results similar to the probabilistic approach (since model error had no time to bias EPS predictions) and significantly better than attractor-based predictions. Indeed, due to generally wider π_{DYN}, the latter are often close to the ignorance point as shown by the histogram of the predictions associated with observed events A in Figure 10. For all lead times, the histogram associated with attractor-based predictions presents a single peak located on the ignorance point. On the contrary, the EPS-based predictions do not show such a behaviour before $t = 5$

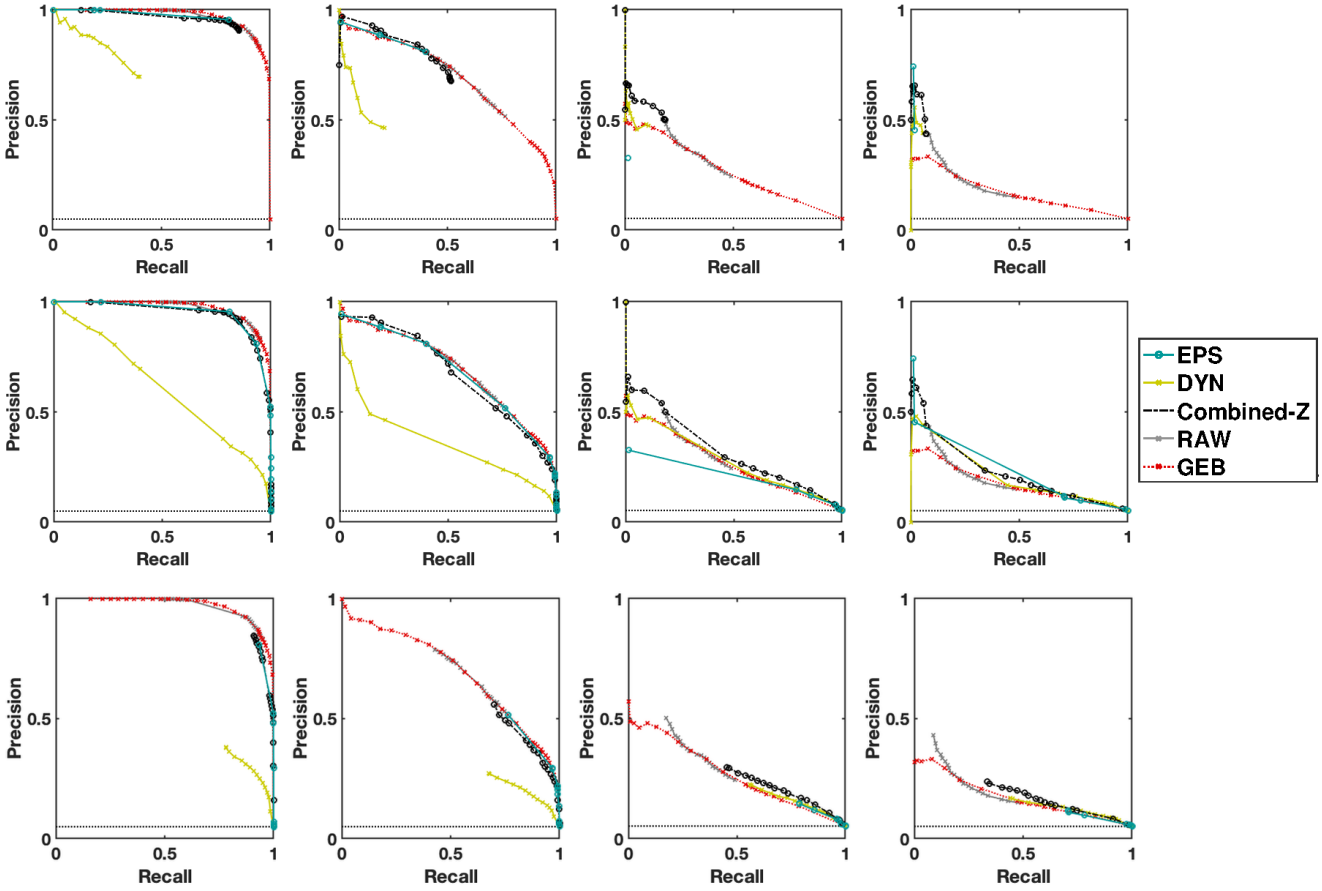


Figure 9. Precision-recall curves showing the predictive skills of possibility distributions EPS, DYN and COMB-Z-9 and probability distributions RAW and GEB for lead times $t = \{1, 3, 5, 7\}$ days (left to right). For the curves associated with the possibilistic approaches, we use $N(A)$, the credibility $0.5\big(N(A) + \Pi(A)\big)$ and $\Pi(A)$ (from top to bottom) as input probabilities.

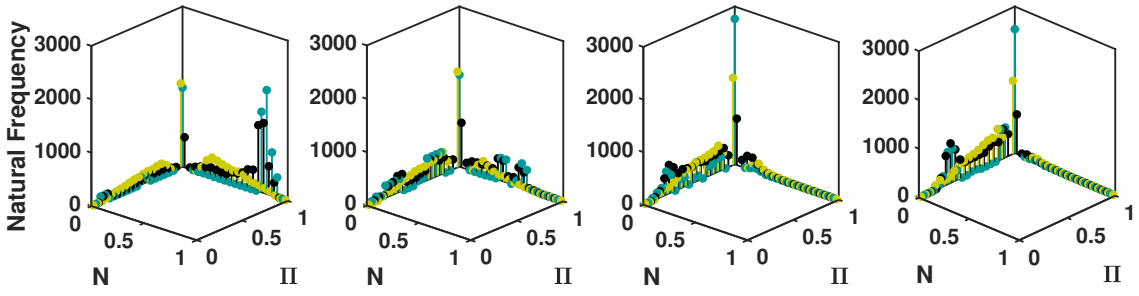


Figure 10. 3-dimensional histograms of the possibilistic predictions associated to verification of A for lead times $t = \{1, 3, 5, 7\}$ days (left to right). Predictions are based on π_{EPS} (blue), π_{DYN} (yellow) or their Z-combination (black).

days. Till that lead time, a large part of the observed events A are associated with a point on the $(\Pi = 1, N > 0)$ axis, meaning tentative acceptance of A. For large lead times, π_{EPS} becomes larger due to the effect of the initial sampling and sensitivity of the model dynamics, both driving ensemble members away from the actual verification with enough time. Combining this distribution to π_{DYN} through the AND operator allows for a narrower final distribution (the peak at the ignorance point of π_{COMB} is smaller in amplitude than the peaks of its components π_{EPS} and π_{DYN}) and provides predictions that discriminate more between A and $\bar{A}$. As shown through the PRC curves, they are also more powerful at large lead times than the predictions given by the probabilistic approach alone, for the same dynamical reasons (model drift, sensibility to ICs).

Using the general aggregation method instead of Zadeh's, do not change significantly the above results. The most notable difference, in favor of the Z-aggregation, is that using the general aggregation restricts even more towards the two extremes (0 and 1) the range of possible recalls.

Practically, using our possibilistic predictor at large lead times and for a given recall, increases the precision by 0.05 for medium recalls and up to 0.3 for small recalls. In other words, for a given hit rate, our framework emits less false alarms, a trend that is all the more marked for small hit rates.

5.3.3 Operational use of the possibilistic concept of ignorance

The information content of a probabilistic prediction $G(O_i)$ of the actual future O_i is evaluated through the ignorance score $S_i = -\log_2 G(O_i)$. The latter characterizes the level of ignorance of the user of such prediction w.r.t. the actual future outcome. On their side, possibilistic frameworks provide predictions in the form of dual measures, the necessity and the possibility of an event, that can be used altogether to characterize the level of ignorance regarding the future outcome to predict, given the evidence at hand. Namely $W = \Pi(A) - N(A)$ is a positive quantity that takes its minimum when $\Pi(A) = N(A) = 0$ ($\bar{A}$ predicted, A being considered impossible) or $\Pi(A) = N(A) = 1$ (A is predicted, $\bar{A}$ being considered impossible) and its maximum when $\big(\Pi(A) = 1, N(A) = 0\big)$ (both A and $\bar{A}$ are possible, none of them is necessary, no tentative acceptance of A or $\bar{A}$ is dictated by the information at hand).

We can consequently wonder: is the probabilistic ignorance S_i (*a posteriori* measured) correlated to the possibilistic level of ignorance W_i (*a priori* measured)? If so, *a priori* observation of the possibilistic level of ignorance could guide for a better use of the probabilistic predictions. Figure 11 aims at answering this question. We compare the Spearman correlation coefficient between the *a posteriori* assessed probabilistic ignorance (for each method, RAW and GEB) and the *a priori* measurable level of possibilistic ignorance (for each possibility distribution, π_{EPS}, π_{DYN} and π_{COMB}). Besides, to highlight results, we

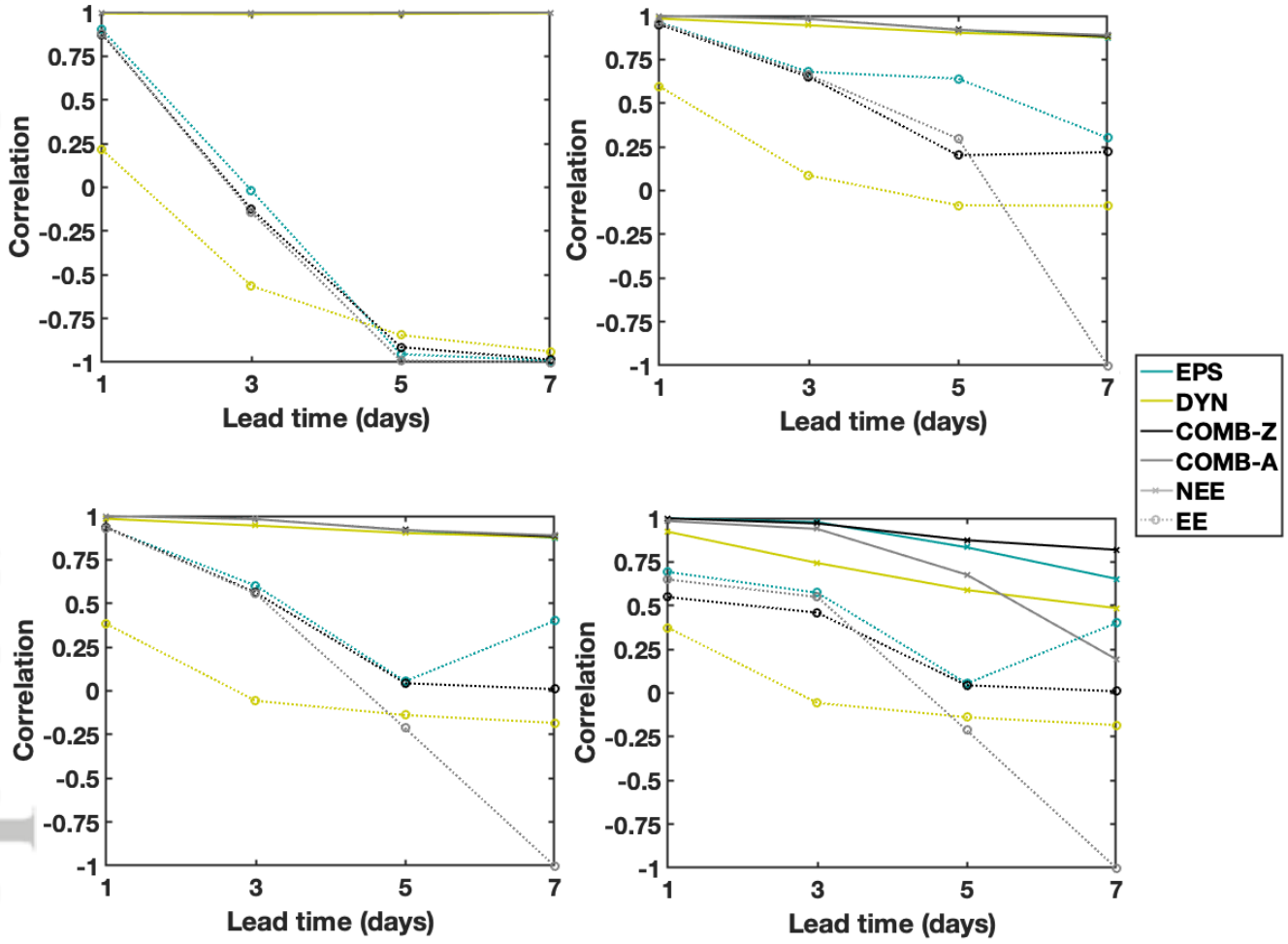


Figure 11. Spearman correlation coefficient between the *a posteriori* propabilistic ignorance score and the level of *a priori* possibilistic ignorance. Results are broken down for EE and NEE, observed (top left) or predicted (others). From left to right and top to bottom, methodologies used are pred-CRED with breakdown of observed EE/NEE, pred-CRED with risk-prone breakdown EE/NEE, pred-CRED with risk-averse breakdown EE/NEE and pred-TENT-AV with risk-averse breakdown EE/NEE.

compare the correlation for observations that belong to the category EE, redefined as "$x \leq q_{0.5} \cup x \geq q_{0.95}$" with the correlation for observations that do not belong to category EE (noted NEE).

Figure 11 reports the correlation between S_i associated with the probabilities derived from possibilistic methodologies and the associated W_i. It would not make sense to directly compare probabilities from GEB or RAW and the possibilistic W_i as the latter are issued from different methodologies.

If we break down results between EE and NEE, we observe that possibilistic (W_i) and probabilistic (S_i) ignorance (reported here in the pred-CRED case) are extremely correlated for NEE, at all lead times and for all possibility distributions. However, in the case of EE, if the correlation is strong and positive at very small lead time (1 day) for COMB-Z, COMB-A and EPS, it becomes strongly negative for lead times above 3 days and all methods. In other words, the level of possibilistic ignorance can be used as a predictor of the information content (i.e. quality) of the pred-CRED prediction only for very small lead times. For larger lead times, the correlation is strong in both EE and NEE case, however of opposite signs which makes it not usable in practice. This pitfall comes from the fact that we break down the correlation results based on the *a priori* unknown future state of the system (EE vs NEE).

What may be more interesting is to break them down w.r.t. the *a priori* known possibilistic prediction, namely: tentative acceptance of EE/A if $N(A) > 0$ (including $\big(N(A) = 0, \Pi(A) = 1\big)$ for the risk-averse option), and tentative acceptance of NEE/$\bar{A}$ if $\Pi(A) < 1$ (including $\big(N(A) = 0, \Pi(A) = 1\big)$ for the risk-prone option).

In the risk-prone version, for tentative acceptance of NEE, the correlation is close to 1 for all possibilistic methods and all lead times, although slightly decreasing with increasing lead times. In other words, when we predict that $\bar{A}$ happens (i.e. $\Pi(A) < 1$ or $\big(N(A) = 0, \Pi(A) = 1\big)$) and associate to it the probability $P(\bar{A}) = \frac{1-\Pi(A)+1-N(A)}{2}$, we get an *a posteriori* probabilistic ignorance that is strongly correlated to the *a priori* possibilistic ignorance W. The latter can consequently be used predictor of the information-content of the possibility-based probability $P(\bar{A})$. The same applies for EE predicted (tentative acceptance of A with associated probability $P(A) = \frac{N(A)+\Pi(A)}{2}$, when $N(A) > 0$) at lead times $t \leq 5$ days for EPS, and lead times $t \leq 3$ days for COMB-Z and COMB-A or lead time $t = 1$ for DYN, all the more than the lead time is small. However, for larger lead times, the correlation coefficient becomes too small to suggest an operational relationship between both types of ignorance. In other words, the possibilistic ignorance for predicted EE is an indicator of the related probabilistic ignorance only for reasonably small lead times, reasonably depending on the method (EPS vs COMB) used. It is interesting to note the case of COMB-A, which provides a strong negative correlation at large lead times. In this case, the larger W, the better the information content of probabilities derived from the possibilistic pred-CRED for predicted EE. This makes sense since larger W generates pred-CRED probabilities that tend towards 0.5 and are consequently less risky that extreme ones.

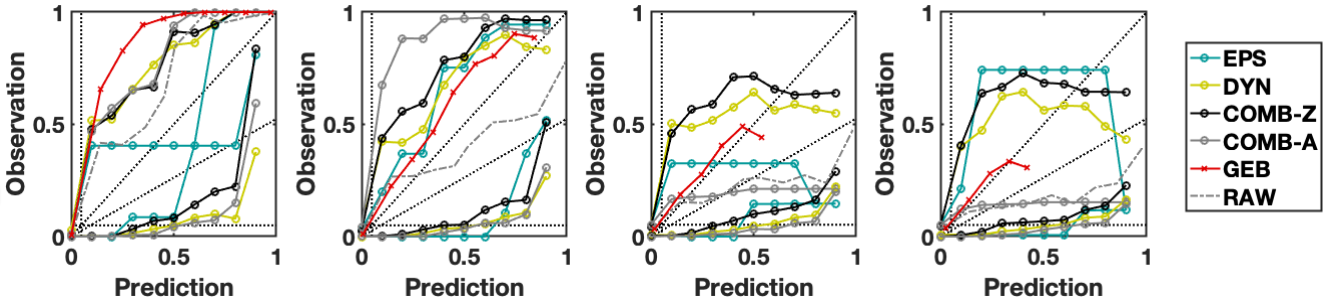


Figure 12. Reliability diagram at lead times $t = \{1, 3, 5, 7\}$ days (left to right). The probabilistic results GEB and RAW are reported in cross-red and dashed grey lines respectively, while the upper and lower bounds of the possibilistic methodologies are in solid-circled lines. Standards elements of comparison are reported in the diagram, namely the diagonal (perfect reliability), the climatological reference (horizontal dotted) and the cone of skill (inside the dashed-dotted secants).

In the risk-averse option, results do not change for NEE predicted: the correlation is still very strong and W_i can be used as a predictor of S_i. When it comes to EE, results are slightly less interesting: beyond 3 days, no possibilistic method shows good positive correlation between W_i and S_i. The former can consequently be used as a predictor of the former only for small lead times, with similar results whatever the possibilistic approach (EPS, DYN, COMB-A, COMB-Z). We observe the same negative correlation for the largest lead time and COMB-A, which has the same interpretation as above.

Finally, we present the correlation observed for probabilities derived, not anymore from pred-CRED but from pred-TENT-AV, in the risk-averse breakdown of predicted EE and NEE. Operationally, results show that only EPS and COMB-Z-9 based methodologies provide W_i and S_i positively correlated at all lead times when NEE are predicted. For predicted EE, a correlation relatively strong (above 0.6) exists for EPS and COMB-A for small lead times, allowing to use to a certain extent W_i as predictor of the information content of S_i. However beyond 3 days, the correlation is too weak to be useful operationally, apart from in the COMB-A case at largest lead time, where we observe again a strong negative correlation.

These results show how and to what extent we can use the full potential of possibilistic measures operationally, that is by deriving equivalent probabilities and by quantifying how informative these are.

5.3.4 Reliability

Figure 12 represents the fuzzy reliability diagram associated with the possibilistic and probabilistic predictions, where lines that are closest to the diagonal show best reliability. For the possibilistic methods, upper and lower bounds of the individual reliability plots obtained by varying $\alpha_i \in [0, 1]$ in $P_{\alpha_i}(A)$ are reported (cf. Section 4.3.4). Both axis are partitioned in 10 bins and we only report the results for bins on the 'Prediction' axis that count at least 10 observations.

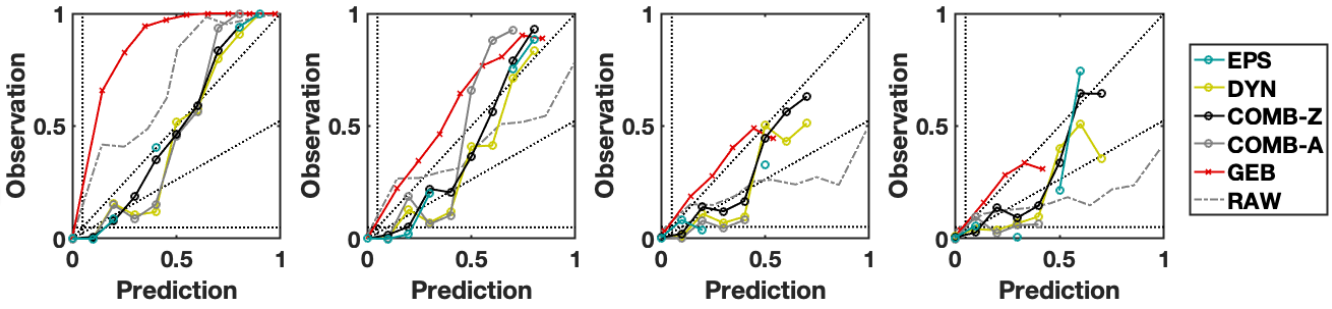


Figure 13. See legend of Figure 12. For possibilistic methodologies, we now extract the credibility and use it as a probability to draw the associated reliability diagrams.

For all lead times, the envelop of the fuzzy reliability plots covers almost the whole range of probability $[0,1]$ while the traditional GEB do not for medium and large lead times. The probabilistic RAW does at all lead times, however the associated reliability diagram falls below the cone of skill beyond lead time 3 days, indicating no resolution. Our approach is consequently capable of providing large probabilities, even for a rare event, without any *a posteriori* recalibration step. Among the different possibilistic approaches, bounds are tighter at small lead times for EPS, however COMB-Z-9 quickly becomes the more interesting methodology for larger lead times. In particular, we note that COMB-A-9 looses resolution beyond 3 days, being not specific enough. For all lead times, at least half of the envelope of the fuzzy reliability plots is contained in the cones of skill, which indicates resolution of the possibility-based probabilities. The perfect reliability line is surrounded by the bounds, apart for probabilities above 0.65 above 5 days.

For a more operational perspective, we analyse the reliability of the possibility-based probabilities derived by means of pred-CRED, that is when we use the credibility as the probabilistic product associated to a possibility distribution. We first note on Figure 13 that the reliability plot is sparse for the EPS method. The latter produces probabilistic predictions focused on the extremes or middle probabilities. The intermediate ones correspond to points falling in the ignorance area, while the upper/lower correspond to peaked distributions towards A or $\bar{A}$. This is all the more visible for short lead times, and experiments show that increasing the archive size N_I allows to reduce the discontinuities. Combining EPS to the more continuous DYN brings continuity in the probabilistic predictions issued from COMB-Z-9. As seen before, we again note that COMB-Z-9 is more informative than both EPS and DYN alone, as it is overall closer to the perfect reliability line than the latter. Finally, in comparison to the GEB approach, COMB-Z-9 is significantly more reliable at small lead times. For larger lead times, it becomes less reliable (namely, overpredictive) than GEB for probabilities below 0.5, however for the upper part of predictive probabilities $(0.5-0.75)$, it is close to perfect reliability while GEB does not output this range of probability at all. COMB-A-9 and RAW produce results similar in essence to the above description of Figure 12.

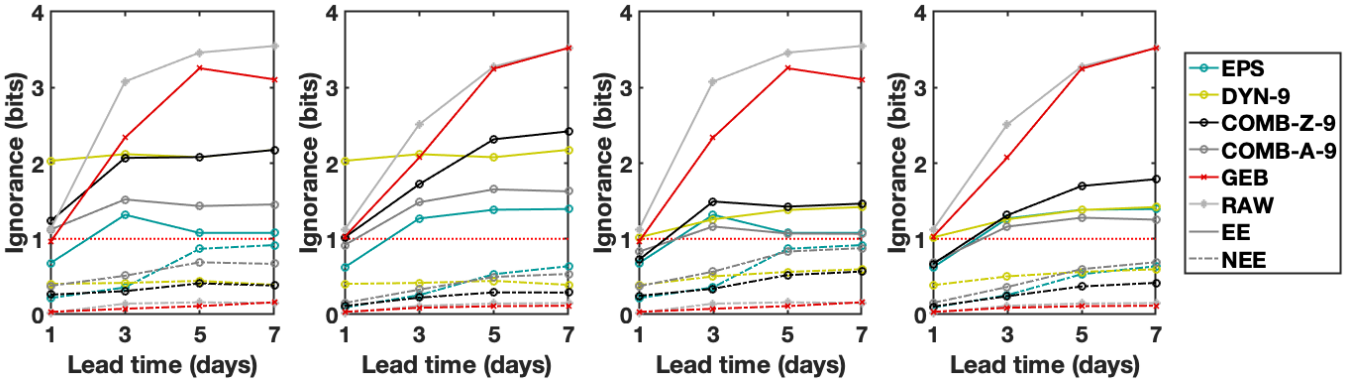


Figure 14. See legend of Figure 7. The method used is pred-CRED. The left two diagrams are based on a time series of length about 6 months and the right two ones on a time series of length about 55 years. Within each block of two, we increase from left to right the EPS archive size from 3 years to 30 years.

5.3.5 Effect of the archive size

We conclude the discussion with a focus on the impact of the archive sizes N_I and N_{I_A} on the predictive performances of our framework. An extended discussion can be found in Le Carrer (n.d.), where we present as well the impact of the size of the archives on the formal guarantees that can be derived. Here, we plot the ignorance score for the following combinations:

- $N_I = 1560$ and $N_{I_A} = 2.10^6$ (the case studied so far: an EPS archive of 30 years and a time series monitoring of the variable of interest of about 55 years) ;
- $N_I = 1560$ and $N_{I_A} = 2.10^4$, that is we lower the time series of the system to less than 6 months ;
- $N_I = 156$ and $N_{I_A} = 2.10^6$, that is we lower the EPS archive to 3 years instead of 30 ;
- $N_I = 156$ and $N_{I_A} = 2.10^4$.

Figure 14 presents the empirical ignorance score similarly to Figure 7, for the possibilistic methodologies EPS, COMB-Z, COMB-A (all in the case of pred-CRED) and the probabilistic GEB and RAW. We observe that increasing the size of the archive $\mathcal{I}_A$ significantly improves beyond 3 days of lead time the information content of the credibility for combined methodologies COMB-Z and COMB-A when it comes to EE. However, again for EE, in both cases the information content of the possibilistic methodologies is above the information content of the probabilistic ones apart for very small lead times for COMB-Z/A, where it is slightly above GEB's. For NEE we observe the opposite effect: increasing the size of the system time series tends to deteriorate slightly performances. Increasing the EPS archive has the opposite effect: it improves the NEE however tends to

deteriorate slightly performances and guarantees for EE. In Le Carrer (n.d.), we develop this counter-intuitive observation and explain how this is due to the limit behaviour of the possibilistic transformation presented in Section 2.1. More points tend to lower the level γ such that $\pi(x) \geq \gamma : \forall x$, that is the minimal possibility degree for any event of interest A: $\Pi(A) \geq \gamma$, in particular the EE we are interested in. Consequently, for possibility profiles that do not show a peak in the area of definition of A (e.g. $\Pi(\bar{A}) = 1 \Rightarrow N(A) = 0$), the credibility $C(A) = \frac{N(A)+\Pi(A)}{2}$ is pulled towards lower values, which provides less informative credibility if A is *a posteriori* observed. This phenomenon plays in favor of NEE who have here a large area of definition. On the contrary, when the time series used for dynamical modelling is increased in size, we observe a significant improvement of the information content of DYN for the prediction of EE at all lead times, while the performance is slightly deteriorated for NEE at larger lead times. DYN possibility distributions are built from a set of analogs, n_A that is fixed in size. Increasing the length of the time series will consequently not impact π_{DYN} the same way it does for π_{EPS}. It will increase the density of analogs among which n_A are extracted. This plays in favor of the EE, which were located in scarce areas of the attractor (with a fixed n_A, potentially less distant analogs will be associated). However when it comes to NEE, we can assume that the same applies against them: close to EE areas, EE analogs are taken into account as analogs and consequently lower $N(\bar{A})$ in the associated possibility distribution. The increase (EE) or decrease (NEE) in information content observed on DYN when the size of the archive increases passes on COMB distributions.

Operationally, we conclude that indeed, and as could be expected, the performance of our possibilistic framework depends on the size of the archives at hand. In any case, when it comes to EE prediction, possibility-based information remain globally much more interesting than the purely probabilistic one, especially at large lead times. The EPS archive does not need to be particularly large, while results significantly improve with a longer system monitoring.

Conclusions

In this paper, we have investigated the benefits of using a framework based on possibility theory for interpreting EPSs, and compared it to the standard probabilistic paradigm in the context of extreme event forecasting. In parallel, we have developed a methodology based on dynamical analogs that integrates dynamical information from a time series of the system to the EPS-based possibilistic framework. The possibilistic framework allows us to combine several incomplete sources of knowledge in a consistent manner, and thus to reduce their respective conservatism. Our framework is more direct than the probabilistic one: we do not try to correct misleading EPS-based probabilities. A possibilistic interpretation directly makes sense, without resorting to additional layers of calibration. Moreover, we are able to reproduce the probabilistic predictive skills (PRC at small lead times) and improve them (PRC to a small extent at large lead times, reliability), especially when it comes to EE

without deteriorating significantly the performances for NEE (information content). Different methodologies were introduced and compared (although not exhaustively for shortness of space), showing how risk-averse and risk-prone users could seize the potential of the dual measures to extract predictive probabilities differently from the traditional credibility. However, it turns out that the latter remains globally the best trade-off when it comes to the quality of predictive performances for EE and NEE simultaneously.

Our framework also reveals the strengths and weaknesses of EPSs: at small lead times, the EPS-based information alone is enough to reproduce probabilistic performances, due to low aggregated model error. At larger lead times, however the latter becomes significant, and makes the EPS-based information not sufficient to provide predictions with resolution. That is where the synergy between EPS-based and dynamical-analog-based information allows us to go beyond standard probabilistic performances. However, it would be interesting to see whether the conclusions obtained on the L96 toy system apply to real-world weather EPS.

We also discussed how to use the full potential of the dual possibilistic measures: to derive predictive probabilities and to estimate *a priori* the trust we can have in their informativeness.

Let us now come back to our initial question: echoing Bröcker and Smith (2008), we wondered whether the probability distribution is the best representation of the valuable information contained in an EPS. Our answer would be that it can be at short lead times, when aggregated model error is low; however there is more predictive information and explanatory power to be gained when switching to an imprecise-probability framework at large lead times. Even at short lead times, our framework showed that it could improve e.g. probabilistic reliability and provide an indicator of how informative is the associated credibility. Among the imprecise-probability settings (e.g. credal sets) we chose possibility theory. Conceptually, especially for end-users and predictions, it indeed seems the most intuitive and adapted in this context.

Appendix A: Masson and Denœux (2006)'s methodology to infer a possibility distribution from empirical data

The methodology of Masson and Denœux (2006) to infer a possibility distribution $\pi(x)$ on the stochastic variable $x \in \mathcal{X}$ for which we have a set S of N_s samples, can be summarized as such:

1. First, bin the x-axis in n bins (or classes) b_i centered in x^i: $B = \{b_i, i = 1, \ldots, n\}$ and note n_i their respective population size.

2. Based on the former histogram, compute the simultaneous confidence intervals for multinomial proportions by means of the Goodman's formulation (Goodman, 1965). The latter, reported in Appendix B, provides multinomial confidence

intervals at level $1-\beta$ for the physical 'true' multinomial probabilities. The formulation being based on asymptotic approximations (see full proof reproduced in Appendix A of Masson and Denœux (2006)), a comparative study by May and Johnson (1997) showed that it requires $n > 2$ and minimal class populations $n_i > 5, i = 1, \ldots, n$ to be reliable. The same authors suggest Sison and Glaz (1995) in the contrary case. Other methodologies like the imprecise Dirichlet model of Walley (1996) can be used however they do not offer the same formal guarantees.

We obtain the set of confidence intervals $[p_i^-, p_i^+]$ associated to each true probability p_i of observing the variable x in bin b_i. In the Goodman case, this set of simultaneous confidence intervals guarantees the overall joint confidence level $1-\beta$.

3. If we denote $\mathcal{P}$ the partial order induced by the intervals $[p_i^-, p_i^+]$, then $(b_i, b_j) \in \mathcal{P} \Leftrightarrow p_i^+ < p_j^-$. Find the set of the compatible permutations $\{\sigma_l, l = 1, \ldots, L\}$, where σ_l is the permutation of the indices $\{1, \ldots, n\}$ associated to $\mathcal{P}$ such that $p^+_{\sigma_l(1)} < p^-_{\sigma_l(2)}, p^+_{\sigma_l(2)} < p^-_{\sigma_l(3)}, \ldots, p^+_{\sigma_l(n-1)} < p^-_{\sigma_l(n)}$ or equivalently $\sigma_l(i) < \sigma_l(j) \Leftrightarrow \left(b_{\sigma(i)}, b_{\sigma(j)}\right) \in \mathcal{P}$. σ is a bijection and the reverse transformation σ^{-1} gives the rank of each class b_i in the list of the probabilities sorted according to the partial order $\mathcal{P}$.

4. For each possible permutation σ_l and each class b_i, solve the following linear program:

$$\pi_i^{\sigma_l} = \max_{p1,\ldots,p_n} \sum_{j|\sigma_l^{-1}(j) \leq \sigma_l^{-1}(i)} p_j \tag{A1}$$

under the constraints

$$\begin{cases} \sum_{k=1}^{K} p_k = 1 \\ p_k^- \leq p_k \leq p_k^+ \quad \forall k \in \{1, \ldots, n\} \\ p_{\sigma_l(1)} \leq p_{\sigma_l(2)} \leq \ldots \leq p_{\sigma_l(n)} \quad . \end{cases} \tag{A2}$$

5. Finally, take the distribution dominating all the distributions π^{σ_l}:

$$\pi_i = \max_{l=1,\ldots,L} \pi_i^{\sigma_l} \; \forall i \in \{1, \ldots, n\}. \tag{A3}$$

Such a procedure allows to compute a possibility distribution $\pi(x)$ that dominates with confidence $1-\beta$ the true probability distribution (i.e. in $100(1-\beta)\%$ of the cases). We present it in its principle and brute-force implementation so that the reader

understands the concepts behind it. Yet, this program is limited to small values of n ($n < 10$), mostly due to the complexity of the algorithm providing the list of of permutations following a partial order (which is $O(L)$, where L is the total number of permutations, with worst-case value $L = n!$). Masson and Denœux (2006) derive a simpler computational algorithm, whose solution is shown to be equivalent to the first one. We refer the interested reader to their paper for a full presentation of the tractable version of the algorithm, that we have implemented in this study.

Appendix B: Goodman (1965)'s formulation

Following the problem and notation introduced in Appendix A, if we note:

$$A = \chi^2(1-\beta/n, 1) + N_s \,, \tag{B1}$$

where $\chi^2(1-\beta/n, 1)$ is the quantile of order $1-\beta/n$ of the chi-square distribution with one degree of freedom, and $N_s = \sum_{i=1}^{n} n_i$ the size of the sample set,

$$B_i = \chi^2(1-\beta/n, 1) + 2n_i \,, \tag{B2}$$

$$C_i = B_i^2 - 4AC_i \,, \tag{B3}$$

$$\Delta_i = \frac{n_i^2}{N_s} \,, \tag{B4}$$

then the bounds of the confidence intervals $[p_i^-, p_i^+]$ associated to the true probabilities p_i of observing the variable x in bin b_i, $i = 1, \ldots, n$ are given by:

$$[p_i^-, p_i^+] = \left[\frac{B_i - \sqrt{\Delta_i}}{2A}, \frac{B_i + \sqrt{\Delta_i}}{2A} \right] . \tag{B5}$$

Author contributions. NLC conceived the presented idea, designed and implemented the research and wrote the article. SF reviewed the article.

Competing interests. The authors declare that they have no conflict of interest.

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 the reviewers and the associate editor who greatly helped to improve the quality of this manuscript. We also thank M. Broccardo for fruitful discussions on PRCs, and N. Berthier for his careful reviewing.

References

Allen, S., Ferro, C. A. and Kwasniok, F. (2019), 'Regime-dependent statistical post-processing of ensemble forecasts', *Quarterly Journal of the Royal Meteorological Society* **145**(725), 3535–3552.

Bröcker, J. and Smith, L. A. (2007), 'Increasing the Reliability of Reliability Diagrams', *Weather and Forecasting* **22**(3), 651–661.

Bröcker, J. and Smith, L. A. (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.

Buizza, R., Milleer, M. and Palmer, T. N. (1999), 'Stochastic representation of model uncertainties in the ECMWF ensemble prediction system', *Quarterly Journal of the Royal Meteorological Society* **125**(560), 2887–2908.

Daoud, A. B., Sauquet, E., Bontron, G., Obled, C. and Lang, M. (2016), 'Daily quantitative precipitation forecasts based on the analogue method: Improvements and application to a French large river basin', *Atmospheric Research* **169**, 147–159.

De Cooman, G. and Aeyels, D. (1999), 'Supremum preserving upper probabilities', *Information Sciences* **118**(1), 173–212.

Delgado, M. and Moral, S. (1987), 'On the concept of possibility-probability consistency', *Fuzzy Sets and Systems* **21**(3), 311–318.

Deyle, E. R. and Sugihara, G. (2011), 'Generalized Theorems for Nonlinear State Space Reconstruction', *PLoS One* **6**(3).

Dubois, D., Foulloy, L., Mauris, G. and Prade, H. (2004), 'Probability-Possibility Transformations, Triangular Fuzzy Sets, and Probabilistic Inequalities', *Reliable computing* **10**(4), 273–297.

Dubois, D. and Prade, H. (1982), 'On several representations of uncertain body of evidence', *Fuzzy Information and Decision Processes* pp. 167–181.

Dubois, D. and Prade, H. (2012), *Possibility theory: an approach to computerized processing of uncertainty*, Springer Science & Business Media.

Dubois, D. and Prade, H. (2015), Possibility theory and its applications: Where do we stand?, *in* 'Springer handbook of computational intelligence', Springer, pp. 31–60.

Dubois, D., Prade, H. and Sandri, S. (1993), On Possibility/Probability Transformations, *in* R. Lowen and M. Roubens, eds, 'Fuzzy Logic: State of the Art', Springer Netherlands, Dordrecht, pp. 103–112.

Friederichs, P. and Hense, A. (2007), 'Statistical Downscaling of Extreme Precipitation Events Using Censored Quantile Regression', *Monthly Weather Review* **135**(6), 2365–2378.

Friederichs, P., Wahl, S. and Buschow, S. (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 Katzfuss, M. (2014), 'Probabilistic forecasting', *Annual Review of Statistics and Its Application* **1**, 125–151.

Gneiting, T., Raftery, A. E., Westveld, A. H. and Goldman, T. (2005), 'Calibrated Probabilistic Forecasting Using Ensemble Model Output Statistics and Minimum CRPS Estimation', *Monthly Weather Review* **133**(5), 1098–1118.

Good, I. J. (1966), 'How to Estimate Probabilities', *IMA Journal of Applied Mathematics* **2**(4), 364–383.

Goodman, L. A. (1965), 'On simultaneous confidence intervals for multinomial proportions', *Technometrics* **7**(2), 247–254.

Graziani, C., Rosner, R., Adams, J. M. and Machete, R. L. (2019), 'Probabilistic Recalibration of Forecasts', *arXiv preprint arXiv:1904.02855* .

Hagedorn, R., Hamill, T. M. and Whitaker, J. S. (2008), 'Probabilistic forecast calibration using ecmwf and gfs ensemble reforecasts. part i: Two-meter temperatures', *Monthly Weather Review* **136**(7), 2608–2619.

Hamill, T. M. and Colucci, S. J. (1997), 'Verification of Eta–RSM Short-Range Ensemble Forecasts', *Monthly Weather Review* **125**(6), 1312–1327.

Hamill, T. M. and Scheuerer, M. (2018), 'Probabilistic precipitation forecast postprocessing using quantile mapping and rank-weighted best-member dressing', *Monthly Weather Review* **146**(12), 4079–4098.

Hamill, T. M. and Whitaker, J. S. (2006), 'Probabilistic Quantitative Precipitation Forecasts Based on Reforecast Analogs: Theory and Application', *Monthly Weather Review* **134**(11), 3209–3229.

Hamill, T. M., Whitaker, J. S. and Wei, X. (2004), 'Ensemble reforecasting: Improving medium-range forecast skill using retrospective forecasts', *Monthly Weather Review* **132**(6), 1434–1447.

Hartmann, D. L., Buizza, R. and Palmer, T. N. (1995), 'Singular Vectors: The Effect of Spatial Scale on Linear Growth of Disturbances', *Journal of the Atmospheric Sciences* **52**(22), 3885–3894.

Hose, D. and Hanss, M. (2019), 'Possibilistic calculus as a conservative counterpart to probabilistic calculus', *Mechanical Systems and Signal Processing* **133**, 106290.

Klir, G. J. (2006), 'Uncertainty and information', *Foundations of Generalized Information Theory* .

Le Carrer, N. (n.d.), 'Possibly extreme, probably not: Is possibility theory the route for risk-averse decision-making?', *Atmospheric Science Letters* p. e01030.

Le Carrer, N. and Green, P. L. (2020), 'A possibilistic interpretation of ensemble forecasts: experiments on the imperfect lorenz 96 system.', *Advances in Science and Research* **17**, 39–39.

Legg, T. P. and Mylne, K. R. (2004), 'Early Warnings of Severe Weather from Ensemble Forecast Information', *Weather and Forecasting* **19**(5), 891–906.

Leith, C. E. (1974), 'Theoretical Skill of Monte Carlo Forecasts', *Monthly Weather Review* **102**(6), 409–418.

Liu, B. (2006), 'A survey of credibility theory', *Fuzzy Optimization and Decision Making* **5**(4), 387–408.

Lorenz, E. N. (1956), 'Empirical orthogonal functions and statistical weather prediction'.

Lorenz, E. N. (1969), 'Atmospheric Predictability as Revealed by Naturally Occurring Analogues', *Journal of the Atmospheric Sciences* **26**(4), 636–646.

Lorenz, E. N. (1996), Predictability: A problem partly solved, *in* 'Proc. Seminar on predictability', Vol. 1.

Ma, J., Yang, M., Han, X. and Li, Z. (2017), 'Ultra-Short-Term Wind Generation Forecast Based on Multivariate Empirical Dynamic Modeling', *IEEE Transactions on Industry Applications* **54**(2), 1029–1038.

Masson, M.-H. and Denœux, T. (2006), 'Inferring a possibility distribution from empirical data', *Fuzzy Sets and Systems* **157**(3), 319–340.

May, W. L. and Johnson, W. D. (1997), 'A sas® macro for constructing simultaneous confidence intervals for multinomial proportions', *Computer methods and Programs in Biomedicine* **53**(3), 153–162.

Mylne, K., Woolcock, C., Denholm-Price, J. and Darvell, R. (2002), Operational calibrated probability forecasts from the ECMWF ensemble prediction system: implementation and verification, *in* 'Preprints of the Symposium on Observations, Data Asimmilation and Probabilistic Prediction', pp. 113–118.

Orrell, D. (2005), 'Ensemble Forecasting in a System with Model Error', *Journal of the Atmospheric Sciences* **62**(5), 1652–1659.

Platzer, P., Yiou, P., Naveau, P., Tandeo, P., Zhen, Y., Ailliot, P. and Filipot, J.-F. (2021), 'Using local dynamics to explain analog forecasting of chaotic systems', *Journal of the Atmospheric Sciences* .

Raftery, A. E., Gneiting, T., Balabdaoui, F. and Polakowski, M. (2005), 'Using Bayesian Model Averaging to Calibrate Forecast Ensembles', *Monthly Weather Review* **133**(5), 1155–1174.

Ramesh, N. and Cane, M. A. (2019), 'The Predictability of Tropical Pacific Decadal Variability: Insights from Attractor Reconstruction', *Journal of the Atmospheric Sciences* **76**(3), 801–819.

Roulston, M. S. and Smith, L. A. (2002), 'Evaluating Probabilistic Forecasts Using Information Theory', *Monthly Weather Review* **130**(6), 1653–1660.

Roulston, M. S. and Smith, L. A. (2003), 'Combining dynamical and statistical ensembles', *Tellus A: Dynamic Meteorology and Oceanography* **55**(1), 16–30.

Saito, T. and Rehmsmeier, M. (2015), 'The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets', *PloS one* **10**(3).

Scheuerer, M. (2014), 'Probabilistic quantitative precipitation forecasting using ensemble model output statistics', *Quarterly Journal of the Royal Meteorological Society* **140**(680), 1086–1096.

Sison, C. P. and Glaz, J. (1995), 'Simultaneous confidence intervals and sample size determination for multinomial proportions', *Journal of the American Statistical Association* **90**(429), 366–369.

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', Springer International Publishing, Cham, pp. 289–296.

Sugihara, G. and May, R. M. (1990), 'Nonlinear forecasting as a way of distinguishing chaos from measurement error in time series', *Nature* **344**(6268), 734–741.

Sugihara, G., May, R., Ye, H., Hsieh, C.-h., Deyle, E., Fogarty, M. and Munch, S. (2012), 'Detecting Causality in Complex Ecosystems', *Science* **338**(6106), 496–500.

Takens, F. (1981), Detecting strange attractors in turbulence, *in* D. Rand and L.-S. Young, eds, 'Dynamical Systems and Turbulence, Warwick 1980', Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 366–381.

Toth, Z. and Kalnay, E. (1997), 'Ensemble Forecasting at NCEP and the Breeding Method', *Monthly Weather Review* **125**(12), 3297–3319.

Trevisan, A. (1995), 'Statistical Properties of Predictability from Atmospheric Analogs and the Existence of Multiple Flow Regimes', *Journal of the Atmospheric Sciences* **52**(20), 3577–3592.

Van den Dool, H. (1994), 'Searching for analogues, how long must we wait?', *Tellus A* **46**(3), 314–324.

Walley, P. (1996), 'Inferences from Multinomial Data: Learning About a Bag of Marbles', *Journal of the Royal Statistical Society: Series B (Methodological)* **58**(1), 3–34.

Wilks, D. S. and Hamill, T. M. (1995), 'Potential Economic Value of Ensemble-Based Surface Weather Forecasts', *Monthly Weather Review* **123**(12), 3565–3575.

Williams, R. M., Ferro, C. A. T. and Kwasniok, F. (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.