# A possibilistic framework for interpreting ensemble predictions in weather forecasting and aggregate imperfect sources of information

*Thesis submitted in accordance with the requirements of the University of Liverpool for the degree of*

## Doctor in Philosophy

*by*

**Noémie Le Carrer**

*Supervised by*

Prof. Scott Ferson and Dr. Peter L. Green

Department of Engineering, Institute for Risk and Uncertainty
University of Liverpool

**1st April 2021**

UNIVERSITY OF LIVERPOOL

Institute for Risk and Uncertainty

# ABSTRACT

Until now, works in the field of tide routing (i.e., optimization of cargo loading and ship scheduling decisions in tidal ports and shallow seas) have omitted the uncertainty of sea level predictions. However, the widely used harmonic tide forecasts are not perfectly reliable. Consequences for the maritime industry are significant: current solutions to tide routing may be made robust through the introduction of arbitrary slack, but they are not optimal. Given the financial implications at stake for every additional centimeter of draft and the catastrophic effects of a grounding, an investigation of tide routing from the perspective of risk analysis is necessary, which we first develop in this PhD thesis.

Predicting future sea level errors w.r.t. tide predictions can be achieved by statistical modelling of these errors, based on historical archives, or by physics-based numerical predictions of these deviations. In the latter option, ensemble forecasting has gained popularity in the field of numerical weather prediction as a way of quantifying the uncertainty on forecasts. Tide-surge ensemble forecasts are thus routinely produced, combining hydrodynamic models with weather ensembles. This type of forecasts is commonly interpreted in a probabilistic way. However, the latter is regularly criticized for not being reliable, especially for predicting extreme events because of the chaotic nature of the dynamics of the atmospheric-ocean system, model error, and the fact that ensemble of forecasts are not, in reality, produced in a probabilistic manner.

In this PhD thesis, we consequently develop an alternative possibilistic framework to interpret and use operationally such ensembles of predictions. In particular, we show by numerical experiments on the Lorenz 96 system that probability theory is not always (e.g. at large lead times and extreme events) the best way to extract the valuable information contained in ensemble predictions. Besides, such a possibilistic perspective eases the combination of different imperfect sources of information about the future state of the system at hand (e.g. dynamical information based on past time series and the analog method), in addition to making more sense without the need of post-processing.

Finally, combining both the scheduling problem and the ensemble interpretation solution, we design a shipping decision model to compute optimal cargo loading and scheduling decisions, given the time series of the fuzzy sea levels in these ports that we derive from a possibilistic interpretation of surge ensemble forecasts. The under keel clearance becomes a possibilistic constraint and the resulting shipping optimization problem is solved by means of an optimisation routine adapted to possibilistic variables. Results obtained on a realistic case study with 7-day-ahead tide surge ensemble predictions are discussed and compared with those given by a probabilistic approach, or by standard practices on ships. After our numerical case studies on the Lorenz 96 system, they illustrate the potential and limitations of a possibilistic interpretation of the weather ensemble forecasts over its probabilistic counterpart in a realistic setting.

**Keywords:** *Robust Optimization*, *Uncertainty modelling*, *Decision making*, *Uncertainty propagation*, *Possibility theory*, *Ensemble predictions*, *Weather forecasting.*

## ACKNOWLEDGEMENT

Although the last year and its social isolation, in particular from the faculty, have clouded the memories attached to this PhD journey, I will do my best to acknowledge all the persons thanks to whom this work came to light.

I am first very thankful to both my supervisors: Prof. Scott Ferson for having jumped on the ship when, while I was on the verge of giving birth to my daughter, I also lost my industrial partner, topic and associated first supervisor. In particular, thank you for the technical freedom you gave me. Conversely, I was very happy to have you by my side, Dr. Peter L. Green, to remind me the rules and habits of academic work.

This work would never have taken this direction nor this shape without the wonderful emulation offered in the corridors and seminar room (to mention these only) of our Institute for Risk and Uncertainty. I thank again its founders for having brought to light such a wonderful place of transdisciplinary knowledge centered around the idea of managing and modelling uncertainty, whatever the field of application. I wish a long life to this intertwining of red bricks and grey matter!

I will in particular remember the women-in-science chats (and much more!) with Caroline, our eclectic discussions between sun and wind with Ander, the comforting hugs from Dominic, our talking walks with Francis, the ever-supporting Diego, the continental memories with Maria and Dominik, the after-work talks with Marco and Tania, the international and stochastic emulation with Matthias, the chit-chat with Conor and Kira, the encyclopedic discussions with Marco B. and all the many more daily interactions that made the salt of the Risk institute and of the physical research world before March 2020.

After a year of work-at-home, I had the pleasure to move to the University Ca'Foscari in Venezia, Italy, where I thank very much the local researchers: Carlo, Fatima, Ivan, Seyum, Gabriele, Marco and their peers, for reminding me the pleasure of lunch conversations. A thought for my long-distance friend Camille and my ever-supportive parents: without your cheering in December, I am sure I would never have found the motivation and energy to revise the core two papers of this PhD.

Last but not least, I dedicate this manuscript to my lovely daughter Shannon. I wish I had time to do some original work with your name sweetie, but the pandemic and institutional responses decided otherwise. Sarà per la prossima volta, te lo assicuro! And finally, infinite thanks go to my latex-guru, Nicolas, whose support for all these years was more than appreciated, in particular with the dreaded above-mentioned language. Long life to pen and paper!

Venezia, 01/04/2021 **Noémie Le Carrer**

Follow the declarations of my supervisors and co-authors giving the permission to use in the PhD the material we published together:

This PhD includes material published with co-authors.

I, Dr. Peter L. Green, give permission for the following papers, co-authored with Noémie Le Carrer, to appear in the PhD thesis:

Noémie Le Carrer, Scott Ferson, and Peter L Green. Optimising cargo loading and ship scheduling in tidal areas. European Journal of Operational Research, 280(3):1082–1094,2020.

Noémie Le Carrer, Scott Ferson, and Peter L Green. Optimising cargo loading and ship scheduling subject to uncertain sea levels. Proceedings of the 8th Workshop on Reliable Engineering Computing, 2018.

Noémie Le Carrer and Peter L Green. A possibilistic interpretation of ensemble forecasts: Experiments on the imperfect Lorenz 96 system. Advances in Science and Research,17:39–39, 2020.

Date: 31/03/2021 Signature:

This PhD includes material published with co-authors.

I, Prof. Scott Ferson, give permission for the following papers, co-authored with Noémie Le Carrer, to appear in the PhD thesis:

Noémie Le Carrer, Scott Ferson, and Peter L Green. Optimising cargo loading and ship scheduling in tidal areas. *European Journal of Operational Research*, 280(3):1082–1094,2020.

Noémie Le Carrer, Scott Ferson, and Peter L Green. Optimising cargo loading and ship scheduling subject to uncertain sea levels. *Proceedings of the 8th Workshop on Reliable Engineering Computing*, 2018.

Noémie Le Carrer and Scott Ferson. Beyond probabilities: A possibilistic framework to interpret ensemble predictions and fuse imperfect sources of information, 2020. Under review.

Date: 30 March 2021 Signature: Scott

# Contents

# Chapter 1

# Introduction

Before getting into the substance of this thesis, I retrace in Section 1.1 its path, which may help the reader to understand its structure as well as the reason why simulated, toy data/systems have been given priority over real data. Elements of the PhD journey are intertwined with the fundamental parts of the literature review devoted to each topic and developed further in the associated chapters (i.e. proper publications). We hope the reader will enjoy the journey as much as we did.

## 1.1 From big data ship routing to possibility theory : Outline of the thesis

This doctoral work starts in September 2016 with a 4-year grant from the EPSRC[1] and ESRC[2] Centre for Doctoral Training in Quantification and Management of Risk and Uncertainty in Complex Systems and Environments. This type of PhD combines one year of full-time 1-year Master of Research (MRes) in Decision-Making under Risk and Uncertainty at the University of Liverpool, followed by 3 years of classical doctoral work at the Institute for Risk and Uncertainty, part of the Department of Engineering of the same university. In such doctoral schemes, the PhD is tied to an industrial partner. Namely, we started to work with Sea Level Research, a local start-up company headed by Dr. Simon Holgate (previously working in the National Oceanography Center, University of Liverpool). The company was developing solutions for maritime companies wanting to optimise their ship scheduling, through improved modelling of actual sea levels.

Thus started a doctoral cycle under the wide topic *Big data adaptative dynamic route planning for high-sea transportation.* The idea was to address an issue of special importance in the process of making the global economy more sustainable : how to green the shipping industry by optimising the fleets' routes ? (see body of work on the topic – e.g. [Davarzani et al., 2016]) Maritime transportation is an activity particularly subject to risk, i.e. the possibility of a loss, due to the complex dynamics and stochastic

[1] Engineering and Physical Sciences Research Council

[2] Economic and Social Research Council

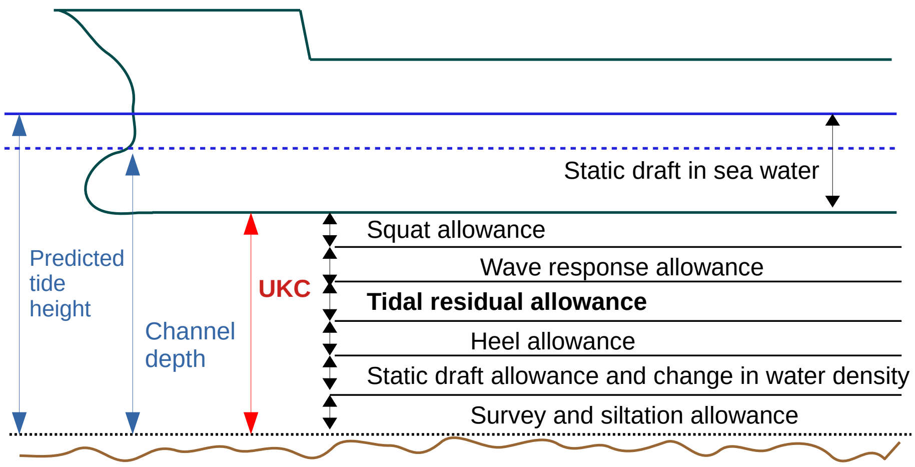
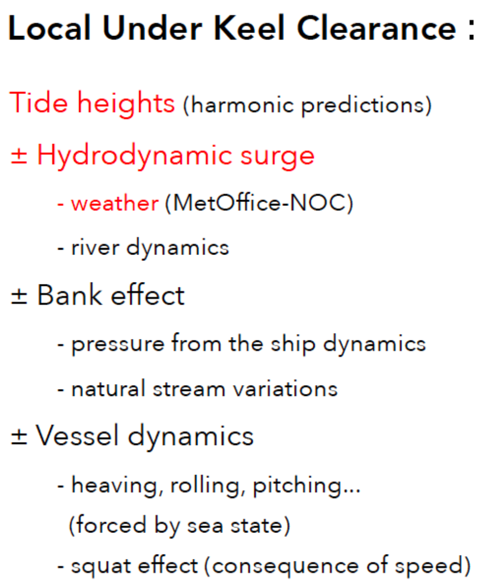


**Figure 1.1** Overview of the under-keel clearance (UKC) bounding problem, where the UKC is the actual water height between the bottom of the hull and the ground : to the water height, we retrieve the ship's nominal static draft and a number of allowances accounting for variations in the actual draft or water height due to a range of factors detailed in Figure 1.2.

nature of the multi-dimensional environment in which it takes place [Goerlandt and Montewka, 2015, Song and Furman, 2013]. From the weather at sea to port variables (berth availability, loading/unloading works), including the volatility of bunker fuel prices or market demand, but also to the less mentioned sea levels in shallow waters, a range of uncertain factors condition the outputs of a shipping operation. In spite of its significant impacts on shipping productivity, the issue of uncertainty has remained marginal in the research on maritime transportation until the last decade. Indeed, as stressed by Song and Furman [2013], due to the complexity and intractability of some shipping problems, authors introduce simplifications (constant speed, single cargo type, basic weather model, etc) that are different from one study to another, making comparison difficult. The introduction of stochasticity in routing is often limited to the modelling of a single or a very limited number of factors, most often the weather [Azaron and Kianfar, 2003], but also market demand [Chuang et al., 2010], or berth occupation and operations [Agra et al., 2015, Yu et al., 2017]. These routing solutions are possibly suboptimal as there are a range of other factors constraining the ships' movements, not taken into account in the problem definition or by means of fixed allowances (see Figure 1.1).

As a result, the general idea of the PhD was to tackle the route-planning problem from a broader viewpoint and first to focus on the analysis of the dependencies of actual routing with environmental features such as the weather in a broader meaning than wind and wave heights (i.e. incorporating sea levels, currents). Then, depending on the amount of satellite vessel tracking data collected during the PhD from collaborations that had to be established, the project would either focus on finding the better way to quantify the link between actual routes and weather features, or on the development



Local Under Keel Clearance :

Tide heights (harmonic predictions)

± Hydrodynamic surge

- weather (MetOffice-NOC)
- river dynamics

± Bank effect

- pressure from the ship dynamics
- natural stream variations

± Vessel dynamics

- heaving, rolling, pitching… (forced by sea state)
- squat effect (consequence of speed)

**Figure 1.2** Factors influencing the UKC. The corresponding allowances are estimated (cf. Figure 1.1) in order to assess a lower bound on the actual UKC at any time. In red, the source of variation of the UKC that will be modelled in this work.

of near optimal, robust and dynamic route-planning algorithms allowing to fuse constantly changing data sources, namely : ship movements, ocean currents, weather data, port information as well as schedules of the shipping lines.

### 1.1.1 Robust maritime shipping optimisation in tidal areas

As a first step to this target, and following the birth of my daughter Shannon in April 2017, I consequently devoted the thesis of the MRes to the study of ship scheduling optimisation subject to uncertain sea levels, with sea level predictions provided by Sea Level Research. Indeed, knowing that an extra centimetre of **draft** (that is the distance between the waterline and the lower point of the ship's hull) corresponds approximately to 50 tons of cargo for an average bulk carrier [Uslu et al., 2017], and that a ton of freight has a value ranging from e.g. US$ 2196 (malting barley) to US$ 223477 (tin), being able to predict accurately water levels in ports translates into significant economic benefits for both shipping operators (economies of scale) and port authorities (vessel throughput).

Deterministic harmonic tide predictions are traditionally used to estimate the future sea levels in shallow waters. From these ones, a shipper can estimate how much freight to load in order to ensure a positive **under-keel clearance** (UKC; the distance between the deepest underwater point of the ship and the seabed – see Figure 1.1), which includes a safety margin dictated from port authorities. Yet sea levels are impacted by environmental factors (wind, pressure, currents) that locally increase or decrease the actual sea levels w.r.t. the harmonic predictions. The difference, hereafter

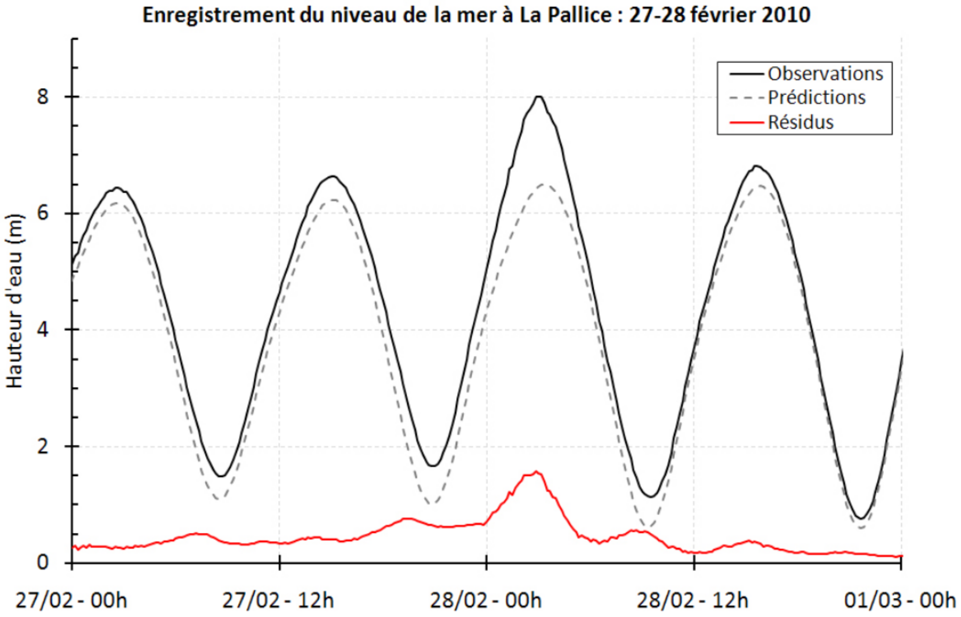


**Figure 1.3** Understanding sea level residuals (in red), that is the difference between tide observations (solid black line) and predictions (dashed line), here observed on 27-28th February 2010 at the La Pallice station, France. Source: *http://tiga.sonel.org*

**residual** (illustrated on Figure 1.3), can be significant. For example, overall British tide stations, residuals are typically 10 cm and rise to 29 cm for high tidal range stations [Flowerdew et al., 2010]. Similarly, sea level residuals can amount to 30% of the total measured sea level in Hillarys Boat Harbour, Western Australia [Makarynskyy et al., 2004]. Whether to load more, depart earlier, or catch a tide window, recent works have shown the economical value of modeling sea level residuals beyond a traditional 'rule-of-the-thumb' safety margin on tide predictions [O'Brien et al., 2002].

Kelareva [2011], Kelareva et al. [2012] first developed the concept of dynamic UKC to optimise ship scheduling and cargo loading decisions of multiple vessels at a single port. To estimate the dynamic UKC, the authors deduct from the port depth and predicted tide, not only the vessel's draft, but also a number of allowances accounting for the dynamical responses of the hull to its environment (squat, heeling, wave, water density variation), the tidal prediction error and the variability of bathymetry [Galor, 2008]. Kelareva [2011] use short-term predictions of the dynamic under-keel clearance provided by the DUKC® software (OMC International, 1993, described in Kelareva et al. [2012], O'Brien et al. [2002]). Specifically, from real-time environmental measurements (water depths, wind, waves, current) and ship information (trim, speed, acceleration), the physical responses to the ship moving in a dynamic environment are computed and the dynamic under-keel clearance is estimated. The optimal cargo loading and short term ship scheduling decisions, given this estimation, are then calculated. Such a solution is based on real-time measurement of the sea state and provides under-keel clearance information for the **upcoming tide-window only** [Kelareva et al., 2012].

The economic gains of such a dynamic modelling of the UKC are documented in a range of case studies, for optimising cargo load and port throughput [O'Brien et al., 2002] or berth-to-berth voyage scheduling optimisation [Hibbert et al., 2019].

Beyond these specific studies, in most of the shipping optimisation problems taking into account tides, water depths are considered as perfectly predictable variables [Xu et al., 2012, Du et al., 2015, Dadashi et al., 2017, Zhen et al., 2017, Yu et al., 2017, Lalla-Ruiz et al., 2016]. When they are not (see [O'Brien et al., 2002] as well as studies on the probabilistic risk assessment of ship grounding in ports [Gucma, 2004, Gucma and Schoeneich, 2008, Abaei et al., 2018]), an allowance, accounting for tide (and possibly bathymetry) prediction error, is introduced. In 2017, to the knowledge of the authors, the modelling of this source of uncertainty was not discussed in the literature. It was consequently worth investigating the robustness and optimality of such modelling, as the introduction of safety margins and/or slack in schedules generally decreases shipping benefits [Kelareva, 2014].

Our first investigations consequently aimed at **filling a gap in the field of ship routing by explicitly considering and modelling the uncertainty in tide predictions on a several-day ahead basis**[3]. A risk analysis of cargo loading and ship scheduling decisions in tidal areas was developed through a realistic case study investigating the research question:

**Q1** How can we optimise the cargo loading and ship scheduling decisions given imperfect tide point forecasts without foregoing safety?

Life is however never short of surprises: our industrial partner Sea Level Research went bankrupt in spring 2017 and consequently decided to put an end to his technical support. The original primary supervisor of this PhD suggested me to change for a topic connected to dynamic recognition and trajectory prediction of military devices, for which he had access to military radar data. However, for ethical reasons (and for efficiency, given my new family responsibilities), I made the choice to continue to investigate the question initiated during the MRes thesis. As a result, Prof. Scott Ferson became my new primary supervisor, along with Dr. Peter L. Green who had followed me as secondary supervisor during the MRes.

Thus, they accompanied me in the preparation and publication of the first journal article of this thesis, *Optimising cargo loading and ship scheduling in tidal areas* [Le Carrer et al., 2020], which extended a conference paper presented at the 8th International Workshop in Reliable Engineering Computing [Le Carrer et al., 2018] and

[3] The DUKC®'s short term UKC predictions are now informed by sea level predictions from two distinct models: a global oceanic "weather" model (coastal currents, mesoscale eddies, etc) and a refined sea level model at the port scale [Uslu et al., 2017]. Both are assimilated by means of a Bayesian recursive approach, where residual are assumed Gaussian, allowing improved 7—day ahead predictions for operational use. These advances were released after our initial investigations.

whose extracts are presented in Appendix A as complement of Chapter 2. Both answer the research question **Q1** when the source of information at hand about future sea levels is **harmonic tide predictions**, along with an archive of actual observations and corresponding tide predictions.

In this article, reported in Chapter 2, through two realistic case studies we show the potential of taking into account the stochastic dimension of sea levels in maritime cargo loading and scheduling decisions, rather than setting a 'rule-of-the-thumb' fixed safety-margin on the ship's draft as it is common use in the field[4]. Results show that the subsequent decision is not only robust in real port and weather conditions, but also closer to optimality. Furthermore, the designed technique remains more interesting in non-stationary settings, namely when sea level variations are artificially increased beyond the extremes of the current residual models.

Precisely, the sea level residuals were modelled very simply through a best-fit (possibly mixture of) parametric distribution(s) fitted to historical time-series of residuals in each port of interest by maximum likelihood optimisation. One way to go would have been to study in depth this probabilistic modelling and refine it by e.g. taking into account the space-time dependence between residuals from different locations and/or time (e.g. by means of a random field approach), in particular its cyclic nature (due to tides), or using alternative approach such as neural networks [Liang et al., 2008, Pashova and Popova, 2011] or superstatistics [Rabassa and Beck, 2015]). Another way would have been to work on the efficient optimisation side of the problem and, instead of a simple Monte-Carlo algorithm nested in a Particle Swarm Optimisation routine, provide a faster optimisation procedure that scales well with the number of ports at hand. A last route consisted in assessing the performance of the optimisation model with another source of sea level information, namely the so-called **ensemble predictions**.

### 1.1.2 Interpreting ensembles of weather predictions: From probability to possibility theory

After a couple of readings on this last trend in the field of weather forecasting, and given my exposure in the Institute of Risk and Uncertainty to alternative theories for modelling uncertainty, I felt rather uncomfortable with the current **all-probabilistic paradigm** [Palmer, 2012, 2017] in ensemble forecasting and especially the way these probabilities were derived. Something did not match between the nature of the dynamical system at hand (atmosphere-ocean), the design of ensembles and their probabilistic interpretation.

[4] Personal correspondence with an officer of the French Merchant Navy

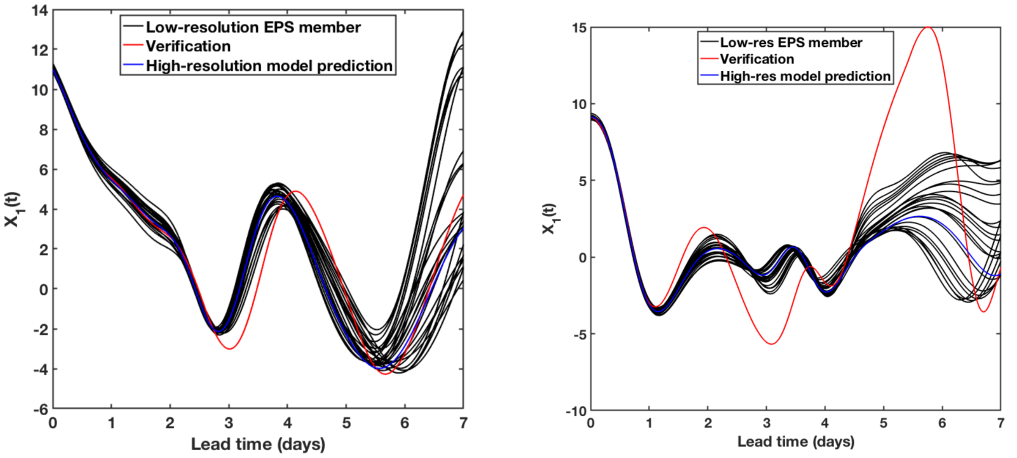


**Figure 1.4** Two EPSs with different initial conditions in the context of the Lorenz 96 system, our experimental test-bed presented in Chapter 3, Section 3.4.1.

#### 1.1.2.1 Ensemble Prediction Systems (EPSs) in weather forecasting

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** (ICs). Limited grid resolution for the 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 (typically 1 to 15 days) **deterministic point predictions**. A shift in paradigm was introduced in parallel of 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 IC estimation, to run the model forward for each of these sampled IC, and to interpret the $M$ resulting predictions in a Monte-Carlo like fashion. Ensemble forecasts (EPSs hereafter) 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) of to directly provide probabilities of observing a given event. See Figure 1.4 for an illustration of EPSs on the Lorenz 96 system, our experimental test-bed introduced in Section 3.4.1.

#### 1.1.2.2 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 (whose state vector's dimension is of the order $10^6$). 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 generally not 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 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**. In other words, on average, the probability derived for a given event does not equal the frequency of verification [Bröcker and Smith, 2007, Smith, 2016, Hamill and Scheuerer, 2018]. Although such probabilistic predictions have higher forecast skill than the **climatology** (that is an history-based probability density of the weather variable at hand), most often they cannot be used as **actionable probabilities**. By design (EPS size limited to $\approx 20 - 50$, 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, Gneiting and Katzfuss, 2014]. This observation is all the more true for **extreme events**, that result from nonlinear interactions at small scales. Such interactions cannot be reproduced in number in a limited-size EPS [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 (see Vannitsem et al. [2020] for an overview). 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.

#### 1.1.2.3 Possibilistic interpretation of ensemble predictions

In view of all this, and especially considering the need to resort to (possibly multiple) calibration steps to provide meaningful probabilistic outputs, we echoed Bröcker and Smith [2008] who questioned the choice of probability distributions as **the best representation of the valuable information contained in an EPS**.

Thus started a few application-orientated tentatives of alternative EPS interpretation. First, and coming back to our shipping problem, the future sea level residual became an interval bounded by the extremes of the corresponding residual EPS (unpublished yet presented at the 2018 Annual Meeting of the European Meteorological Society under the title *Robust optimisation of cargo loading and ship scheduling in tidal areas* [Le Carrer, 2018a]). The interval being overconfident (as noted in the above literature review), we fitted a logistic regression by means of 'carefully chosen predictors' (namely mean amplitude of the surge and width of the EPS) to be able to predict whether the observation would fall into or outside the interval (EPS) bounds. The overall success rate being rather good (91%), we then turned the probability to be in/above or below (the dreaded case) into a fuzzy safety margin, that we finally introduced into our shipping optimisation algorithm. The resulting decision model was introduced in a wider port simulation and results were presented from the perspective of the port authorities (Can individual shipping optimisation improve port traffic or do we need central intervention?) on the occasion of the 2018 Conference "Mathematics Applied in Transport and Traffic Systems" [Le Carrer, 2018b].

Such modelling approaches were interesting and promising technically yet not really satisfying when it came to their interpretation and to the question of "**making**

**sense**". We criticized the probabilistic approach for not making sense and needing postprocessing and we ended up building similar layers of modelling followed by error correction.

That's how **possibility theory** became quickly appealing for its intuitive rationale, its explanatory power w.r.t. our weather application and for its potential in terms of uncertainty communication in the field of forecasting. We started to wonder if this *"weaker theory than probability [...] also relevant in non-probabilistic settings where additivity no longer makes sense"* [Dubois et al., 2004] could provide an interesting alternative, in a context where conceptual and practical limitations **restrict the applicability of a density-based (i.e. additive) interpretation of EPSs**.

A first possibilistic ensemble dressing (for the parallel with the probabilistic ensemble dressing of Roulston and Smith [2003], that consisted in "dressing" each ensemble member with historical error statistics) was designed and, for lack of data or connection with the weather forecasting research community, tested on a the Lorenz 96 toy system [Lorenz, 1996] (L96), commonly used for such studies on EPSs [Wilks, 2006, Williams et al., 2014]. Results were presented at the Annual Meeting of the European Meteorological Society in September 2019, which led to the proceedings [Le Carrer and Green, 2020] whose framework and results are reported in Appendix B. Although novel and promising in the case of extreme event predictions, this model was a first parametric try which suffered from serious 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.

We consequently extended it into a purely data-driven model which made more sense. The model is introduced and studied in two journal articles: 1) *Possibly Extreme, Probably Not: Is possibility theory the route for risk-averse decision-making?* (accepted in Atmospheric Science Letters), reported in Chapter 3, where we investigate the **guarantees** that can be derived from such a possibilistic interpretation of EPSs and compare them to the traditional probabilistic postprocessing, thus focusing on the **continuous** interpretation of possibility distributions ; and 2) *Beyond probabilities: A possibilistic framework to interpret ensemble predictions and fuse imperfect sources of information* (in review at the Quarterly Journal of the Royal Meteorological Society after minor revisions in May 2021), reported in Chapter 4, where we show how to combine **dynamical information** extracted from a time series of the system, to the pure EPS interpretation by means of possibility theory and we focus on the **binary** analysis of the resulting possibility distributions, on the **information content** of possibilistic and probabilistic interpretations respectively, as well as on the operationability of our approach when it comes to weather forecasting. In both cases, given the nature of the issue with the probabilistic treatment of EPSs, we pay a particular attention to the case of extreme event predictions.

In these two works we investigate the following research questions:

**Q2a** 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?

**Q2b** Can we simultaneously maintain or improve the prediction skills compared to those of standard probabilistic interpretations?

**Q2c** How can we combine such a possibilistic framework with insights about the local dynamics of the system?

**Q2d** Can a possibilistic treatment of the EPS provide more formal guarantees than a probabilistic interpretation? If yes, at what cost?

**Q2e** Can we operationally use the possibilistic outputs at their full potential, that is more than simply deriving associated probabilities?

### 1.1.3 Application to shipping optimisation subject to uncertain sea levels

Results being promising, we closed the loop by applying our possibilistic framework to the ship scheduling problem designed in Chapter 2, now investigating the research questions:

**Q3a** How valuable is the information extracted from the storm-surge EPS, either via a probabilistic approach or via a possibilistic approach, for an application such as maritime shipping optimisation?

**Q3b** In particular, is this information more valuable for this specific application than a classical Monte-Carlo-based optimisation using harmonic tide predictions and historical best-fit modelling of sea level residuals in each port?

With the limitation of having at hand very few EPS data regarding sea levels (1 year over 2 British ports), we adapted a procedure designed in Hose et al. [2018] to deal with global optimisation when input parameters or constraints are possibilistic (here, sea level predictions) and compared the practical performances of the various modelling of sea level uncertainty, that is a) no modelling yet fixed "rule-of-the-thumb" safety margin on the ship's draft ; b) history-based, stationary best-fit probabilistic modelling of tide residuals; and c) possibilistic and probabilistic interpretations of residual EPSs in each port of call. We were thus able to draw conclusions as regards the relative potential of each methodology when it comes to practical use in the shipping industry, as well as to confront our EPS-based methodology with true data, although limited in number – hence the limitation in our conclusions. This last study, *A possibilistic* *interpretation of ensemble predictions: Application to shipping optimisation in tidal areas*, is reported in Chapter 5 and is presented at the 2021 edition of the European Safety and Reliability Conference.

### 1.1.4 The closure problem

We are aware that this PhD work could be pursued in many directions, the most important being:

(i) Pursuing the investigations on the information content (w.r.t. predictability) and optimal use for prediction (e.g. depending on the attitude towards risk of the end-user) of both necessity and possibility measures, started in Chapter 4 ;

(ii) Studying the potential of the possibilistic interpretation of EPSs in clearly non-stationary systems ;

(iii) Gathering more data to analyse the asymptotic behavior (when the problem dimension increases) of the possibility-based shipping optimisation procedure.

However, due to the pandemic situation and its impact on daily work schedules (in particular nursery and university closure), we leave these to future works. This PhD thesis can be seen as a first investigation of the potential of possibility theory in the field of the prediction of nonlinear dynamical systems, with special interest in weather forecasting and contribution in optimisation-based applications of possibilistic predictions.

## 1.2 Structure of the thesis and summary of original contributions

In this section, we summarize the PhD structure developed and justified in the previous section, and we report the original contributions associated to each chapter.

We start with Chapter 2, namely the journal paper *Optimising cargo loading and ship scheduling in tidal areas* [Le Carrer et al., 2020] accepted in 2019 in the European Journal of Operational Research. In this publication, a shipping decision model is created, which consists in the optimisation of the predicted net shipping benefit given information on the future sea levels of each port of call. By means of two realistic case studies involving a small bulk carrier, 2 and 3 British ports respectively, a farm commodity, and the harmonic tide predictions along with an history of residuals in each port as sources of information on future sea levels, we show therein that:

- Trusting the harmonic tide predictions blindly leads to potentially dramatic loss;
- Adding an *a priori* fixed safety margin allows the method to become robust, however it becomes potentially suboptimal. Besides, assessing the optimal

safety margin is a problem-dependent task that requires a number of simulations beforehand ;

- Taking into account the stochastic nature of the sea level residuals by means of a best-fit modelling in each port of call allows to provide optimally robust decisions, that is decisions that are robust yet closer to optimality (the latter being defined as the decisions that would have been made in the presence of perfect information on the future sea levels) ;
- The stochastic decision model is robust to unseen extreme sea level variations, accounting for (limited) non-stationarity of the underlying residual distributions.

Overall, the novelty of this work is a **risk analysis** of cargo loading and ship scheduling decisions in tidal areas and a method to provide optimally robust decisions. It aims at raising awareness of the economic potential of taking into account sea level uncertainty in scheduling decisions more finely than a 'rule-of-the-thumb' safety margin, not only for the more studied expensive freight and large ships (e.g. [O'Brien et al., 2002]) but also for the masses of small vessels (mini-bulkers), cheap commodities (grains) and small ports strongly affected by tidal effects (i.e. with limited dredging), which in the current context of transportation greening may be a non-negligible lever of progress.

The latter article is an extension of the conference paper *Optimising cargo loading and ship scheduling subject to uncertain sea levels* Le Carrer et al. [2018], presented at the 8th Workshop for Reliable Engineering Computing in 2019. We present it in Annex A as it contains a comparison and discussion of the performances of a range of classical objective functions (also said risk metric: 'mean regret', 'mean risk', 'chance constrained', 'worst case') used in our decision model.

Chapter 3 follows by introducing the possibilistic interpretation of EPSs developed during our thesis: *Possibly Extreme, Probably Not: Is possibility theory the route for risk-averse decision-making?*, published in the journal Atmospheric Science Letters in January 2021 [Le Carrer, 2021]. It follows a conference paper presenting an earlier parametric possibilistic ensemble dressing, *A possibilistic interpretation of ensemble forecasts: Experiments on the imperfect Lorenz 96 system* [Le Carrer and Green, 2020] on the occasion of the 2019 Annual Meeting of the European Meteorological Society.

In this article, by means of numerical experiments on an imperfect version of the Lorenz 96 system, we investigate the formal guarantees associated to our approach and compare them empirically to those provided by a classical probabilistic interpretation of EPSs, in the case of both extreme and non-extreme events. Our contributions are the following:

- To the knowledge of the author, the first interpretation of EPSs by means of possibility theory;
- The analysis of the resulting predictive possibility distributions in the continuous perspective, namely the analysis of the formal guarantees associated with the confidence intervals that could be derived on future predictions;
- Their empirical comparison to the reliability of a classical probability-based interpretation. We show that the confidence intervals based our methodology overpass the latter in two cases: 1) at very small lead times for both common and extreme events, where they are as reliable yet narrower; 2) more blatantly, at intermediate and large lead times for extreme events, where they remain guaranteed and can be brought close to perfect reliability even for particularly rare events, yet at the expense of precision. In particular, we raise the potential of such an interpretation for risk-averse end-users.

Chapter 3 suffers from two limitations: 1) by focusing on the continuous reading of predictive distributions, we do not exploit the full potential of the possibilistic concepts of necessity, possibility and ignorance; and 2) we do not take into account the local dynamics of the system (initial conditions), which makes EPS-based predictive distributions rather conservative. Chapter 4, *Beyond probabilities: A possibilistic framework to interpret ensemble predictions and fuse imperfect sources of information*, submitted at the Quarterly Journal of the Royal Meteorological Society in July 2020 (minor revisions submitted in May 2021), addresses them both.

Our contributions therein are the following:

- The investigation of the benefits of using our possibilistic framework for interpreting EPSs, in the case of binary predictions. Predictive skills are assessed by means of the ignorance score, PRC curves and reliability diagrams with the credibility used in place of a classical probability, and compared to those of a traditional probability-based interpretation. The possibilistic approach performs at least as well as the probabilistic treatment, and overpasses it for large lead times and extreme events.
- The introduction and comparison of alternative methodologies to use the dual measures necessity/possibility at their full potential instead of the traditional credibility, when it comes to make a binary prediction. We show the potential for predictions better tailored to the end-user needs (e.g. risk-averse, risk-prone).
- The development of a methodology based on dynamical analogs to model in a possibilistic way the information about the future state of the system extracted from a time series recording of the system.

- A discussion on the options for combining the EPS-based and dynamical-based incomplete sources of information about the future state of the system.
- The physical interpretation of our framework: transfer and synergy of information between the two incomplete sources of information at different levels according to the lead time, and a development around the concept of ignorance.
- An investigation of how and when to use the dual possibilistic measures to derive a predictive probability and estimate *a priori* how much we can trust them.

Finally, Chapter 5, combines the two major works of this PhD together by using EPSs to make sea level residuals predictions in the shipping optimisation problem. *A possibilistic interpretation of ensemble predictions: Application to shipping optimisation in tidal areas*, that will be presented at the ESREL 2021 conference, thus gathers the following contributions:

- The design of a methodology inspired from Hose et al. [2018] to run our shipping decision model with possibilistic sea levels as inputs ;
- The application of the possibilistic treatment of EPSs to a small data set (1 year) of residual predictions in two ports, and the assessment of their prediction skill w.r.t. a classical probabilistic interpretation ;
- The comparison of the robustness and optimality of the shipping optimisation procedures according to the source of sea level information (tides and archive of residual, EPSs) and their treatment (deterministic, possibilistic, probabilistic).

Overall, given the small size of our dataset, we observe that the probabilistic best-fit modelling of residuals in each port is the best way to go for shipping companies. However, the possibilistic treatment may become competitive for small-size problems (limited number of ports) and larger (sea levels observations+EPS) archives.

Finally, Chapter 6 provides a global conclusion and discussion of the results and contributions of this PhD and draws the main directions for future works.

As a last note, we mention the existence of a side-PhD work about efficient global optimisation, not presented here for coherence, yet published as a short paper under the title *Robust efficient global optimisation via adaptive surrogate refinement* in the Proceedings in Applied Mathematics and Mechanics 2019 [Le Carrer et al., 2019].

## Bibliography

Mohammad Mahdi Abaei, Ehsan Arzaghi, Rouzbeh Abbassi, Vikram Garaniya, Mohammadreza Javanmardi, and Shuhong Chai. Dynamic reliability assessment of ship grounding using bayesian inference. *Ocean Engineering*, 159:47–55, 2018.

Agostinho Agra, Marielle Christiansen, Alexandrino Delgado, and Lars Magnus Hvattum. A maritime inventory routing problem with stochastic sailing and port times. *Computers and Operations Research*, 61:18–30, 2015.

Amir Azaron and Farhad Kianfar. Dynamic shortest path in stochastic dynamic networks: Ship routing problem. *European Journal of Operational Research*, 144(1): 138–156, 2003.

Jochen Bröcker and Leonard A. Smith. Increasing the Reliability of Reliability Diagrams. *Weather and Forecasting*, 22(3):651–661, 2007. doi: 10.1175/WAF993.1.

Jochen Bröcker and Leonard A. Smith. From ensemble forecasts to predictive distribution functions. *Tellus A: Dynamic Meteorology and Oceanography*, 60(4):663–678, 2008. doi: 10.1111/j.1600-0870.2007.00333.x.

Roberto Buizza. Ensemble Forecasting and the Need for Calibration. In *Statistical Postprocessing of Ensemble Forecasts*, pages 15–48. Elsevier, 2018. ISBN 978-0-12-812372-0. doi: 10.1016/B978-0-12-812372-0.00002-9.

Roberto Buizza, M Milleer, and Tim N Palmer. Stochastic representation of model uncertainties in the ECMWF ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society*, 125(560):2887–2908, 1999. doi: 10.1002/qj.49712556006.

Tzung-Nan Chuang, Chia-Tzu Lin, Jung-Yuan Kung, and Ming-Da Lin. Planning the route of container ships: A fuzzy genetic approach. *Expert Systems with Applications*, 37(4):2948–2956, 2010.

Ali Dadashi, Maxim A Dulebenets, Mihalis M Golias, and Abdolreza Sheikholeslami. A novel continuous berth scheduling model at multiple marine container terminals with tidal considerations. *Maritime Business Review*, 2(2):142–157, 2017.

Hoda Davarzani, Behnam Fahimnia, Michael Bell, and Joseph Sarkis. Greening ports and maritime logistics: A review. *Transportation Research Part D: Transport and Environment*, 48:473–487, 2016.

Yuquan Du, Qiushuang Chen, Jasmine Siu Lee Lam, Ya Xu, and Jin Xin Cao. Modeling the impacts of tides and the virtual arrival policy in berth allocation. *Transportation Science*, 49(4):939–956, 2015.

Didier Dubois, Laurent Foulloy, Gilles Mauris, and Henri Prade. Probability-Possibility Transformations, Triangular Fuzzy Sets, and Probabilistic Inequalities. *Reliable computing*, 10(4):273–297, 2004. doi: 10.1023/B:REOM.0000032115.22510.b5.

Jonathan Flowerdew, Kevin Horsburgh, Chris Wilson, and Ken Mylne. Development and evaluation of an ensemble forecasting system for coastal storm surges. *Quarterly Journal of the Royal Meteorological Society*, 136(651):1444–1456, 2010.

P. Friederichs and A. Hense. Statistical Downscaling of Extreme Precipitation Events Using Censored Quantile Regression. *Monthly Weather Review*, 135(6):2365–2378, 2007. doi: 10.1175/MWR3403.1.

Petra Friederichs, Sabrina Wahl, and Sebastian Buschow. Postprocessing for Extreme Events. In Stéphane Vannitsem, Daniel S. Wilks, and Jakob W. Messner, editors, *Statistical Postprocessing of Ensemble Forecasts*, pages 127–154. Elsevier, 2018. ISBN 978-0-12-812372-0. doi: https://doi.org/10.1016/B978-0-12-812372-0.00005-4.

Wiesław Galor. Determination of dynamic under keel clearance of maneuvering ship. *Journal of KONBiN*, 8(1):53–60, 2008.

Tilmann Gneiting and Matthias Katzfuss. Probabilistic forecasting. *Annual Review of Statistics and Its Application*, 1:125–151, 2014.

Tilmann Gneiting, Adrian E. Raftery, Anton H. Westveld, and Tom Goldman. Calibrated Probabilistic Forecasting Using Ensemble Model Output Statistics and Minimum CRPS Estimation. *Monthly Weather Review*, 133(5):1098–1118, 2005. doi: 10.1175/MWR2904.1.

Floris Goerlandt and Jakub Montewka. Maritime transportation risk analysis: review and analysis in light of some foundational issues. *Reliability Engineering and System Safety*, 138:115–134, 2015.

Carlo Graziani, Robert Rosner, Jennifer M Adams, and Reason L Machete. Probabilistic Recalibration of Forecasts. *arXiv preprint arXiv:1904.02855*, 2019.

L Gucma and M Schoeneich. Probabilistic model of underkeel clearance in decision making process of port captain. *TransNav: International Journal on Marine Navigation and Safety of Sea Transportation*, 2(2), 2008.

Lucjan Gucma. Risk based decision model for maximal ship entry to the ports. In *Probabilistic Safety Assessment and Management*, pages 3022–3027. Springer, 2004.

Thomas M Hamill and Stephen J Colucci. Verification of Eta–RSM Short-Range Ensemble Forecasts. *Monthly Weather Review*, 125(6):1312–1327, 1997.

Thomas M. Hamill and Michael Scheuerer. Probabilistic precipitation forecast postprocessing using quantile mapping and rank-weighted best-member dressing. *Monthly Weather Review*, 146(12):4079–4098, 2018. doi: 10.1175/MWR-D-18-0147.1.

D. L. Hartmann, R. Buizza, and T. N. Palmer. Singular Vectors: The Effect of Spatial Scale on Linear Growth of Disturbances. *Journal of the Atmospheric Sciences*, 52(22): 3885–3894, 1995. doi: 10.1175/1520-0469(1995)052<3885:SVTEOS>2.0.CO;2.

Gregory Hibbert, David O'Brien, et al. Berth to berth voyage schedule optimisation-a torres strait case study. In *Australasian Coasts and Ports 2019 Conference: Future directions from 40 [degrees] S and beyond, Hobart, 10-13 September 2019*, page 569. Engineers Australia, 2019.

Dominik Hose, Markus Mäck, and Michael Hanss. A possibilistic approach to the optimization of uncertain systems. *1*, 2018.

Elena Kelareva. The "DUKC Optimiser" ship scheduling system. In *2011 International Conference on Automated Planning and Scheduling System Demonstrations*, 2011.

Elena Kelareva. *Ship Scheduling with Time-Varying Draft Restrictions: A Case Study in Optimisation with Time-Varying Costs*. PhD thesis, The Australian National University, 2014.

Elena Kelareva, Sebastian Brand, Philip Kilby, Sylvie Thiebaux, and Mark Wallace. CP and MIP Methods for Ship Scheduling with Time-Varying Draft. In *Twenty-Second International Conference on Automated Planning and Scheduling*, ICAPS '12, 2012.

Eduardo Lalla-Ruiz, Christopher Expósito-Izquierdo, Belén Melián-Batista, and J Marcos Moreno-Vega. A set-partitioning-based model for the berth allocation problem under time-dependent limitations. *European Journal of Operational Research*, 250(3): 1001–1012, 2016.

N Le Carrer, S Ferson, and P. L Green. Optimising cargo loading and ship scheduling subject to uncertain sea levels. 8th Workshop on Reliable Engineering Computing, 2018.

Noemie Le Carrer. Robust optimisation of cargo loading and ship scheduling in tidal areas. Annual Meeting of the European Meteorological Society, 2018a.

Noemie Le Carrer. Optimising ship scheduling subject to uncertain sea levels: Application to port traffic. `https://d2k0ddhflgrk1i.cloudfront.net/CiTG/Over%20faculteit/Afdelingen/Transport%20%26%20Planning/Conferences/Matts/Le%20Carrer.pdf`, 2018b. Conference Mathematics Applied in Transport and Traffic Systems.

Noémie Le Carrer. Possibly extreme, probably not: Is possibility theory the route for risk-averse decision-making? *Atmospheric Science Letters*, page e01030, 2021.

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

Noémie Le Carrer, David Moens, and Matthias Faes. Robust efficient global optimisation via adaptive surrogate refinement. *PAMM*, 19(1):e201900474, 2019.

Noémie Le Carrer, Scott Ferson, and Peter L Green. Optimising cargo loading and ship scheduling in tidal areas. *European Journal of Operational Research*, 280(3):1082–1094, 2020.

T. P. Legg and K. R. Mylne. Early Warnings of Severe Weather from Ensemble Forecast Information. *Weather and Forecasting*, 19(5):891–906, 2004. doi: 10.1175/1520-0434(2004)019<0891:EWOSWF>2.0.CO;2.

C. E. Leith. Theoretical Skill of Monte Carlo Forecasts. *Monthly Weather Review*, 102 (6):409–418, 1974. doi: 10.1175/1520-0493(1974)102<0409:TSOMCF>2.0.CO;2.

SX Liang, MC Li, and ZC Sun. Prediction models for tidal level including strong meteorologic effects using a neural network. *Ocean Engineering*, 35(7):666–675, 2008.

Edward N Lorenz. Predictability: A problem partly solved. In *Proc. Seminar on predictability*, volume 1, 1996.

Oleg Makarynskyy, D Makarynska, Michael Kuhn, and WE Featherstone. Predicting sea level variations with artificial neural networks at hillarys boat harbour, western australia. *Estuarine, Coastal and Shelf Science*, 61(2):351–360, 2004.

K Mylne, C Woolcock, J Denholm-Price, and R Darvell. 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*, pages 113–118, 2002.

Terry O'Brien et al. Experience using dynamic underkeel clearance systems: selected case studies and recent developments. In *30th PIANC-AIPCN Congress 2002*, page 1793. Institution of Engineers, 2002.

David Orrell. Ensemble Forecasting in a System with Model Error. *Journal of the Atmospheric Sciences*, 62(5):1652–1659, 2005. doi: 10.1175/JAS3406.1.

Tim Palmer. The primacy of doubt: Evolution of numerical weather prediction from determinism to probability. *Journal of Advances in Modeling Earth Systems*, 9(2): 730–734, 2017.

TN Palmer. Towards the probabilistic earth-system simulator: A vision for the future of climate and weather prediction. *Quarterly Journal of the Royal Meteorological Society*, 138(665):841–861, 2012.

Lyubka Pashova and Silviya Popova. Daily sea level forecast at tide gauge burgas, bulgaria using artificial neural networks. *Journal of sea research*, 66(2):154–161, 2011.

Pau Rabassa and Christian Beck. Superstatistical analysis of sea-level fluctuations. *Physica A: Statistical Mechanics and its Applications*, 417:18–28, 2015.

Adrian E. Raftery, Tilmann Gneiting, Fadoua Balabdaoui, and Michael Polakowski. Using Bayesian Model Averaging to Calibrate Forecast Ensembles. *Monthly Weather Review*, 133(5):1155–1174, 2005. doi: 10.1175/MWR2906.1.

Mark S. Roulston and Leonard A. Smith. Combining dynamical and statistical ensembles. *Tellus A: Dynamic Meteorology and Oceanography*, 55(1):16–30, 2003. doi: 10.3402/tellusa.v55i1.12082.

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

Leonard A. Smith. Integrating Information, Misinformation and Desire: Improved Weather-Risk Management for the Energy Sector. In Philip J. Aston, Anthony J. Mulholland, and Katherine M.M. Tant, editors, *UK Success Stories in Industrial Mathematics*, pages 289–296. Springer International Publishing, Cham, 2016. ISBN 978-3-319-25454-8. doi: 10.1007/978-3-319-25454-8_37.

Jin-Hwa Song and Kevin C Furman. A maritime inventory routing problem: Practical approach. *Computers and Operations Research*, 40(3):657–665, 2013.

Zoltan Toth and Eugenia Kalnay. Ensemble Forecasting at NCEP and the Breeding Method. *Monthly Weather Review*, 125(12):3297–3319, 1997. doi: 10.1175/1520-0493(1997)125<3297:EFANAT>2.0.CO;2.

Burak Uslu, Andy Taylor, Greg Hibbert, Rafael Soutelino, et al. Connecting sea level forecasts with the bulk export industry. *Australasian Coasts and Ports 2017: Working With Nature*, page 1084, 2017.

Stéphane Vannitsem, John Bjørnar Bremnes, Jonathan Demaeyer, Gavin R Evans, Jonathan Flowerdew, Stephan Hemri, Sebastian Lerch, Nigel Roberts, Susanne Theis, Aitor Atencia, et al. Statistical postprocessing for weather forecasts–review, challenges and avenues in a big data world. *Bulletin of the American Meteorological Society*, pages 1–44, 2020.

Daniel S Wilks. Comparison of ensemble-mos methods in the lorenz'96 setting. *Meteorological Applications*, 13(3):243–256, 2006.

Daniel S. Wilks and Thomas M. Hamill. Potential Economic Value of Ensemble-Based Surface Weather Forecasts. *Monthly Weather Review*, 123(12):3565–3575, 1995. doi: 10.1175/1520-0493(1995)123<3565:PEVOEB>2.0.CO;2.

R. M. Williams, C. A. T. Ferro, and F. Kwasniok. A comparison of ensemble post-processing methods for extreme events. *Quarterly Journal of the Royal Meteorological Society*, 140(680):1112–1120, 2014. doi: 10.1002/qj.2198.

Dongsheng Xu, Chung-Lun Li, and Joseph Y-T Leung. Berth allocation with time-dependent physical limitations on vessels. *European Journal of Operational Research*, 216(1):47–56, 2012.

Shucheng Yu, Shuaian Wang, and Lu Zhen. Quay crane scheduling problem with considering tidal impact and fuel consumption. *Flexible Services and Manufacturing Journal*, 29(3-4):345–368, 2017.

Lu Zhen, Zhe Liang, Dan Zhuge, Loo Hay Lee, and Ek Peng Chew. Daily berth planning in a tidal port with channel flow control. *Transportation Research Part B: Methodological*, 106:193–217, 2017.

## Chapter 2

# Optimising cargo loading and ship scheduling in tidal areas

This chapter consists in the article published in 2019 in the *European Journal of Operational Research* [Le Carrer et al., 2020]. The contributions of each author are the following: NLC conceived the presented idea, found the data, designed and implemented the research and then wrote the article, that SF and PLG reviewed.

Therein we aim at filling a gap in the field of ship routing by explicitly considering and modelling the uncertainty in tide predictions on a several-day ahead basis, and thus addressing the question of the optimal draft allowance accounting for tide residuals in tidal areas. That safety margin accounts for variations between harmonic tide predictions and actual sea levels. It is generally fixed in a "rule-of-the-thumb" manner (operationally) or barely considered (in research works). Our study was consequently, to the knowledge of the authors in 2018 (see 2021 update in Section 1.1.1 of Chapter 1), the first to describe a stochastic methodology to model several days ahead sea level residuals and take them into account in the global ship scheduling and cargo loading optimisation process. We develop a "poor man's" probabilistic approach to account for the uncertainty on sea levels in each port, where the latter are considered as stochastic variables and classical distributions are fitted to them by means of a maximum-likelihood approach applied on archives of historical observations (cheap to acquire, hence the "poor man" expression). This methodology is applied to two realistic case studies of maritime shipping between British ports and allows to show the potential of a stochastic-based perspective on sea level residuals in maritime shipping problems in tidal areas.

We add in Appendix A extracts of the earlier conference paper [Le Carrer et al., 2018] that complement this work with a more in-depth presentation of the objective functions that one can design to evaluate the risk associated to a decision in a stochastic context. Different common objective functions are presented and their performance are compared in a similar case study as the first one developed in the next article (see Section 2.5.1).

Finally, we add to the original article the Figures 2.1 and 2.2 for easing the understanding of the concepts and procedures at hand.

## Optimising cargo loading and ship scheduling in tidal areas

N. Le Carrer, S. Ferson and P. L. Green

*European Journal of Operational Research* (accepted August 2019)

**Abstract**

This paper describes a framework that combines decision theory and stochastic optimisation techniques to address tide routing (i.e. optimisation of cargo loading and ship scheduling decisions in tidal ports and shallow seas). Unlike weather routing, tidal routing has been little investigated so far, especially from the perspective of risk analysis. Considering the journey of a bulk carrier between $N$ ports, a shipping decision model is designed to compute cargo loading and scheduling decisions, given the time series of the sea level point forecasts in these ports. Two procedures based on particle swarm optimisation and Monte Carlo simulations are used to solve the shipping net benefit constrained optimisation problem. The outputs of probabilistic risk minimisation are compared with those of net benefit maximisation, the latter including the possibility of a 'rule-of-the-thumb' safety margin. Distributional robustness is discussed as well, with respect to the modelling of sea level residuals. Our technique is assessed on two realistic case studies in British ports. Results show that the decision taking into account the stochastic dimension of sea levels is not only robust in real port and weather conditions, but also closer to optimality than standard practices using a fixed safety margin. Furthermore, it is shown that the proposed technique remains more interesting when sea level variations are artificially increased beyond the extremes of the current residual models.

## 2.1 Introduction and literature review

### 2.1.1 Ship scheduling in tidal areas

A ship's draft is the distance between the waterline and the bottom of the hull. It is a fundamental characteristic of a ship and forms a major constraint in terms of scheduling or cargo loading decisions because a poor choice can lead to grounding in tidal areas or shallow waters. Yet the research on ship loading has mostly focused on operations safety and logistic aspects (see for instance a review in Christiansen et al. [2007]). The question of scheduling with time-varying draft was not tackled until recently, when Kelareva and colleagues developed a deterministic procedure to optimise ship scheduling and cargo loading decisions of multiple vessels at a single port [Kelareva, 2011, Kelareva et al., 2012]. Their procedure is based on the increasingly popular concept of dynamic under-keel clearance.

The under-keel clearance is the distance between the deepest underwater point of the ship and the seabed. In the traditional static approach, the under-keel clearance is computed as the difference between the water depth (combining channel depth and tide prediction) and the nominal ship draft. The objective is then to maintain an under-keel clearance at least equal to either a given minimum value or a given percentage of the ship draft, depending on port policy. On the contrary, the dynamic approach deducts from the channel depth and predicted tide, not only the nominal draft, but also a number of allowances accounting for the dynamical responses of the hull to its environment (squat, heeling, wave, water density variation), the tidal prediction error and the variability of bathymetry [Galor, 2008]. Kelareva [2011] use short-term predictions of the dynamic under-keel clearance provided by the DUKC® software (OMC International, 1993, described in Kelareva et al. [2012], O'Brien et al. [2002]). Specifically, from real-time environmental measurements (water depths, wind, waves, current) and ship information (trim, speed, acceleration), the physical responses to the ship moving in a dynamic environment are computed and the dynamic under-keel clearance is estimated. The optimal cargo loading and short term ship scheduling decisions, given this estimation, are then computed. [O'Brien et al., 2002] report two case studies showing the added value of using such a dynamic under-keel clearance approach in port operations, for both shippers (freight savings and increase in export value) and port operators (reduced dredging costs in the long term, increased ship departure/arrival windows and consequently reduced congestion, contribute scientific knowledge to estimation of the minimal under-keel clearance).

Such a solution is based on real-time measurement of the sea state and provides under keel clearance information for the upcoming tide-window only [Kelareva et al., 2012]. Being deterministic, safety margins have to be introduced as the under-keel clearance is only estimated *a priori*. One can ask whether taking into account the stochastic nature of sea levels (and, consequently, the under-keel clearance) could reduce this safety margin to some theoretical minimum - this is one of the aspects investigated in the current paper. Besides, the planification horizon allowed by the procedure described above is relatively short (one tide) - the current work addresses relatively longer time scales.

The work of Kelareva et al. [2012] was extended to a shipping cost optimisation problem for a fleet considering time-varying draft restrictions at waypoints, variable ship speed and cargo loads as well as flow control through busy waterways [Kelareva, 2014]. The specific waterway ship scheduling problem was later formulated by Lalla-Ruiz et al. [2016], who integrated tide as a constraint in their approach to optimally schedule the flow of incoming and outgoing ships through different shipping channels (so that the waiting times were globally minimised).

Similarly, researchers focusing on the berth allocation problem, which aims at scheduling berth and crane allocation to optimise port throughput, introduced tide as a constraint only quite recently. While early works [Xu et al., 2012, Du et al., 2015] were more concerned with the quantification of the economic impact of tides on port operations, recent studies developed practical models and solutions for berth scheduling optimisation [Dadashi et al., 2017, Zhen et al., 2017] or quay crane allocation [Yu et al., 2017] in tidal ports.

### 2.1.2 Shipping optimisation in stochastic environments

Maritime transportation is an activity particularly subject to risk, i.e. the possibility of a loss, due to the complex dynamics and stochastic nature of the multi-dimensional environment in which it takes place. From the weather at sea to port variables (berth availability, loading/unloading works), including the volatility of bunker fuel prices, a range of uncertain factors condition the outputs of a shipping operation. In spite of its significant impacts on shipping productivity, the issue of uncertainty has remained marginal in the research on maritime transportation until recently. Indeed, as stressed by Song and Furman [2013], due to the complexity and intractability of some shipping problems, authors introduce simplifications (constant speed, single cargo type, basic weather model, etc) that are different from one study to another, making comparison difficult. The introduction of stochasticity is often limited to the modelling of a single or a very limited number of factors (e.g. weather [Azaron and Kianfar, 2003], market demand [Chuang et al., 2010], weather and berth occupation [Agra et al., 2015]).

Water depth is also an uncertain factor that should not be neglected. Although tide forecasts used to predict the water depths in shallow seas are traditionally given by harmonic analysis from past observations, a range of causes can modulate the observed water levels. These encompass weather influence, river discharge, the interaction between currents, shallow water seabed and ship traffic [NOREL, 2014] and lead to significant deviations between astronomical tides and actual water level observations (called residuals hereafter: the difference between observations and predictions). Flowerdew et al. [2010] estimate that the root mean square error on the high tide predictions in UK tide stations is typically 10 cm and rises to 29 cm for high tidal range stations. Makarynskyy et al. [2004] note that sea level residuals can amount to 30% of the total measured sea level in Hillarys Boat Harbour, Western Australia.

The uncertainty about future water depths has a significant impact on shipping optimisation. First, as shown in the case study presented in Section 2.2.1, even for a small-sized carrier of horizontal dimensions $85$ m $\times 15$ m, one additional centimetre of under-keel clearance can be turned into an extra freight of $13.05$ metric tons (mt) whose value ranges from US\$ $2,556$ for a single hold of malting barley [Agriculture and Horticulture Development Board, 2017] to US\$ $223,477$ for a single hold of tin [Quandl, 2017] with little increase of operational costs in short journeys[1]. Secondly, when it costs thousands of dollars a day to operate the same vessel, missing a tide-window because of a negative anomaly in the water depth is significantly costly to the shipper, to say nothing about the cost of grounding and its potential environmental consequences. Another economic justification is found in O'Brien et al. [2002], who shows how the use of the DUKC software (deterministic dynamic under-keel clearance estimation) allowed 123 vessels to load an additional 743,246 tonnes of coal (an average of 6,042 tonnes per vessel) in the Port of Hay Point, Australia, in the 1996/1997 financial year. Resulting freight savings amounted to approximately US$7,500,000 and the increase in export earnings summed up to US$30,000,000.

### 2.1.3 Robustness in shipping optimisation

In most of the approaches mentioned in Section 2.1.1, water depths are considered as perfectly predictable variables. When they are not (see [O'Brien et al., 2002] as well as studies on the probabilistic risk assessment of ship grounding in ports [Gucma, 2004, Gucma and Schoeneich, 2008, Abaei et al., 2018]), an allowance, accounting for tide (and possibly bathymetry) prediction error, is introduced. To the knowledge of the authors, the modelling of this source of uncertainty is not discussed in the literature. It is consequently worth investigating the robustness and optimality of such modelling, as the introduction of safety margins always decreases shipping benefit [Kelareva, 2014].

Although Kelareva et al. [2012] introduced a conservative 15-minute departure window for each departure/arrival decision, the authors justified the slack as a way to take into account the inertia of large ships in port operations rather than to account for sea level uncertainties. The large operational costs of ships tend to prevent the shippers from adding significant slack in their schedule [Christiansen et al., 2007], as a ship is only productive when it is sailing. Again, it is worth investigating the robustness of a dynamic under-keel clearance-based shipping optimisation with respect to real port conditions, namely delays.

An original approach to robustness in ship routing and scheduling is finally found in Christiansen and Fagerholt [2002], who introduced the concept of risky arrival. A penalty cost proportional to the risk of a given schedule is integrated in the optimisation procedure of the transportation cost of a fleet. The work of Brown et al. [1997] should

[1] To paraphrase Kelareva [2014]: if the fuel consumption of the empty ship is 20% less than that of the laden ship [Endresen et al., 2004] and if the increase in fuel consumption is linear in draft difference, then if the vessel at hand shows a difference of 2.47m between laden and empty draft, 1 cm of extra draft equals to 0.08% of fuel consumption increase. For short sea shipping journeys and small extra load draft, this can be neglected. In bad weather, such an assumption might hold only for very small extra loads as fuel consumption is sensitive to both weather and ship draft (both increasing the friction resistance), i.e. load [Bertram, 2012]. Ship characteristics would then be needed to assess the actual added-value of loading more in a rough sea.

also be mentioned as it questions the applicability of mathematical optimisation in a real port context. The authors especially highlight the situation where a small change in the model inputs leads to a radically different optimal solution. The concept of persistence is introduced as a new feature of the optimisation model so that small changes in the input values do not drastically change the nature of the optimal solution.

### 2.1.4 Objective and contribution

The present work aims at filling a gap in the field of ship routing by explicitly considering and modelling the uncertainty in tide prediction. A robust analysis of cargo loading and ship scheduling decisions in tidal areas is drawn through a realistic case study. The question at hand is: how can we optimise the cargo loading and ship scheduling decisions given imperfect sea level forecasts without foregoing safety?

To this purpose, a maritime shipping decision model is introduced. The model assumes that an industrial operator has sea level forecasts at $N$ ports, at a given time $t_0$, over a prediction horizon $T$. On this basis, the operator has to decide the total amount of a given commodity to load at the first port, and the fraction of this cargo that will be delivered in each of the remaining ports, as well as the estimated arrival and departure times in each port. The deliveries all have to satisfy the constraints of the inventory routing problem, namely to match a given demand in each port. We assume that all ports have unlimited storage capacities.

Our model computes the ‘optimal’ solution to such a problem by taking into account the uncertainty in actual sea levels w.r.t. tide predictions. However, the reader must note that our framework does not address the uncertainty associated with the under-keel clearance arising from dynamical responses to the sea state (heeling, heaving, squat effect), nor from the bathymetry. We limit our approach to uncertainty about still water levels resulting from deviations to the tide predictions. These additional dynamical sources of uncertainty could still be integrated to a similar approach in order to address the open water problem (see for instance Briggs et al. [2013] for empirical methods to estimate the squat allowance, Quy et al. [2007] to quantify the ship response to waves and Drwięga et al. [2017] to assess the heel components).

In the following, our model allows us to demonstrate the economic potential of a robust under-keel clearance optimisation. Beyond the application to industrial shipping, for which the bulk cargo load is quite flexible, this work wants to raise awareness of the economic potential for small vessels (mini-bulkers), cheap commodities (grains) and small ports strongly affected by tidal effects (i.e. limited dredging). In the current context of transportation greening [Davarzani et al., 2016], we expect this to be an important area for future applications.

The reader must keep in mind that, to clearly demonstrate the potential of the proposed tide routing approach, we deliberately omit the uncertainties associated with weather, as well as berth availability and cargo handling capacity. With the increase of slow steaming practices [Mallidis et al., 2018], weather does not currently represent a serious limitation since generous journey times are planned and ships are no longer expected to be at the maximum of their performance. The authors aim to further study the limitation represented by berth congestion in future works. For many of the small ports that the current work targets, neglecting the aforementioned uncertainties does not represent a significant issue, as the overall ship and cargo flow is not at its full capacity. For the larger and busier ports, this is indeed a question to ask: Is the added value gained from pure tide routing lost in the variability associated with berthing and cargo handling? Or does tide routing helps to smooth port traffic? This is a subject of future work.

Section 2.1 has introduced the motivations for the investigation of robust cargo loading and scheduling optimisation in tidal areas and outlined the state of the art around this issue. Section 2.2 presents the case studies and sets up the shipping decision model. In Section 2.3, the uncertainty on port sea level forecasts is discussed and a robust alternative to the deterministic decision-making process is presented. Section 2.4 describes the implementation as well as the modelling of sea level uncertainty. Section 2.5 discusses the results of our approach, compared to standard techniques based on the results from two realistic case studies. Finally the findings are summarised in Section 2.6 and perspectives are opened.

## 2.2 Shipping decision model

### 2.2.1 Case study

To illustrate the approach in this paper, a case study is presented, which gives the reader context for the model development that is detailed later. We consider a farm cooperative that owns a small-size bulk carrier. The company uses it to carry various farm commodities between ports along the British coast, especially along the route: Liverpool-Portsmouth-Lowestoft.

Given a freight unit value of US\$ $195.61$ per metric ton (for a malting barley freight [Agriculture and Horticulture Development Board, 2017]), 1 cm of additional draft equals an extra freight of $13.05$ mt on the first vessel which conveys an extra profit of US\$ 2,556 (case study 1, cf. parameters in Table 2.1). As described before, although a heavier ship will consume more fuel, for small vessels and short sea voyages, it remains much more profitable for the company to increase the overall cargo loading if possible.

We assume that at time $t_0$, given the demand constraints on commodity $X$ in the delivery ports, the cooperative has to decide the total amount of $X$ to load in the departure port, as well as departure scheduling in each port. To this purpose, the company uses the long term harmonic tide forecasts as sea level forecasts. Indeed, the more recent and accurate models are not available in all ports. Besides, traditional tide forecasts remain the main source of water level information for many shippers. In the following, we use 'decision' to refer to this set of loading/scheduling decisions and 'benefit' to refer to the net benefit resulting from the implementation of the said decision in actual conditions.

Hence the problem of interest: given the economic, vessel and port parameters (Table 2.1), given the information on available quays and maintained depths (Table 2.2), given the tide predictions in all ports[2]:

1. What is the optimal decision, if the harmonic tide forecasts were considered as perfect (from now on called the 'standard approach')?
2. Is this decision robust to actual port and sea level conditions?
3. What is an optimal *and* robust decision if the uncertainty on tide forecasts is taken into account?
4. What shipping benefit can be guaranteed, given a predefined level of acceptable error, from the robust solution?
5. How do the robust solution and guaranteed benefit depend on the model of the tide residuals?
6. Is the procedure robust to unseen (i.e. extreme) sea level variations?

To answer these questions, a first case study is implemented for $t_0 = 13/01/2017-07:30:00$ and $N = 2$ ports. We compute the optimal solution according to our approach and compare it w.r.t. the standard one. We assess its distributional robustness as well. A second larger-scale analysis is then analysed: 175 different $t_0$ are considered, from July 2016 to December 2016. We compare the performance of our model's decision with that of a standard approach, in terms of daily benefit and robustness across this range of $t_0$ values.

### 2.2.2 Model overview

The model used here is a simplified representation of the maritime inventory routing problem. A material is produced at a given rate in a production site (called the loading port) and consumed at other sites (called unloading ports), at specified rates. Given storage capacities in the production and consumption locations, what is the optimal design of routes and fleet schedule that minimises the shipping costs (sailing and port costs) without interrupting any of the production or the consumption in the

[2] Data provided by the UK Environmental Agency.

**Table 2.1** Model parameters

| Type | Param. | Description | Value | | Unit |
|---|---|---|---|---|---|
| | | | Case study 1 | Case study 2 | |
| Journey | $l$ | Mean distance between departure and arrival ports | 440 | {440 , 200} | Nautical miles |
| | $\rho$ | Mean sea water density | 1,250 | | Kilogram per cubic meter |
| Ship design | $v$ | Mean operational sailing speed | 13 | | Knot |
| | $S$ | Ship horizontal surface | 15 × 85 | 25 × 130 | Meter×Meter |
| | $m_{min}$ | Minimum cargo load (ballast) | 1,870 | 3,000 | Metric ton |
| | $m_{max}$ | Deadweight tonnage (carrying capacity) | 5,170 | 25,000 | Metric ton |
| | $r_{50}$ | Half-laden ship draft | 5.2 | 8 | Meter |
| | $f_s$ | Fuel consumption rate of the laden ship at sea | 8 | 11 | Ton per day |
| | $f_p$ | Fuel consumption rate of the ship at port | 1 | 2 | Ton per day |
| Monetary | $C_f$ | Fuel cost | 387 | | US$ per ton |
| | $C_u$ | Other operational costs (staff, maintenance) | 2,500 | | US$ per day |
| | $C_c$ | Average bulk cargo value | 195.6 | | US$ per ton |
| | $C_{bp*}$ | Berthing and loading/ unloading operation cost within normal opening times | 1,239 | 1,486 | US$ per hour |
| | $C_{bp*}$ | Berthing and loading/ unloading operation cost outside of normal opening times | 1,548 | 1,858 | US$ per hour |
| | $C_p$ | Daily port fee | 1,115 | 1,363 | US$ per day |
| Port | $u_p$ | Bulk material (un)loading rate | 1,200 | 1,000 | Ton per hour |
| | | Normal port opening time | $[7:00, 19:00]$ in all ports | | - |
| | $\alpha$ | Minimum allowed under-keel clearance to navigate in port still waters | 10% static draft | | - |
| Forecast | $\Delta t$ | Sea level forecast time step | 15 | | Minute |
| | $T$ | Horizon of the sea level predictions | 3 | 6 | Day |
| Industrial | $a_j$ | Minimal delivery in port $j > 1$ | - | {4,000 , 2,000} | Day |

**Table 2.2** Quay parameters.

| Port | **Maintained quay depth** (Meters) | |
|---|---|---|
| | Case study 1 | Case study 2 |
| Liverpool | 12 | 12 |
| Portsmouth | 3 | 8 |
| Lowestoft | - | 8 |

aforementioned sites? The optimisation is made on an industrial shipping basis. In other words, the shipper owns the material to be shipped and wants to maximise the net benefit of the shipment (the value of the cargo loaded minus the shipping costs). The fleet consists of a single bulk carrier or general cargo ship and the study is restricted to the $N-1$ legs of length $l_j$, $j = \{1, ..., N-1\}$ between the loading (departure, $p_1$) and unloading (arrival, $p_N$) ports with a constant ship speed, $v$, provided by the ship specifications. From this, the goal is to optimise the decision vector $\boldsymbol{d}$ consisting of the departure time $t_j$ and the cargo $m_j$ shipped from each port $p_j$ given the overall tide predictions available at time $t_0$ spanning the horizon $T$ in the entrance channels of all ports, given constraints on the demand $a_j$ in each port and given constraints from the ship design (carrying capacity), safety at sea (minimum acceptable water under keel), port management (opening times and price bands for port labour). For now, unlimited storage capacities in all ports are assumed. The question of rate of production in the departure port (i.e. offer) is not taken into account.

The ship is assumed to be in the departure port at time $t_0$ with empty tanks and the most recent predictions $\hat{X}_j(t)$ at the shipper's disposal for the sea levels in all ports $p_j$, over the horizon $T$. Time is discretized with the time step $\Delta t$ (following the precision in the sea level prediction and observation time series). Here and in the following, in order to simplify the notations, $t_j$ will be relative to the origin of our time axis $t_0$.

### 2.2.3 Model description

The model takes time series of sea level point-forecasts in all ports of call as inputs. Given contextual parameters regarding the journey, including ship characteristics, freight and port management, generic constraints about acceptable under-keel clearance, latest arrival time and cargo load, demand constraints in delivery ports and, finally, the net return computation rule for a journey, it computes the optimal cargo loading and departure time by means of a particle swarm optimisation (PSO) solver. Figure 2.1 provides a graphical overview of the model in the case $N = 2$ ports. This generalises to $N > 2$.

#### 2.2.3.1 Journey parameters

Table 2.1 defines the model's input parameters. A few comments and justifications are provided here.

The operational speed $v$ is assumed to be fixed and constant over the journey (as it is often the case in maritime shipping models). Operational port costs are subject to price bands. Although most often docks and loading / unloading operations are accessible 24 hours a day 7 days a week, the cost of such operations depends on the local port schedule, e.g. midweek vs weekend periods for Liverpool port [Peel Ports

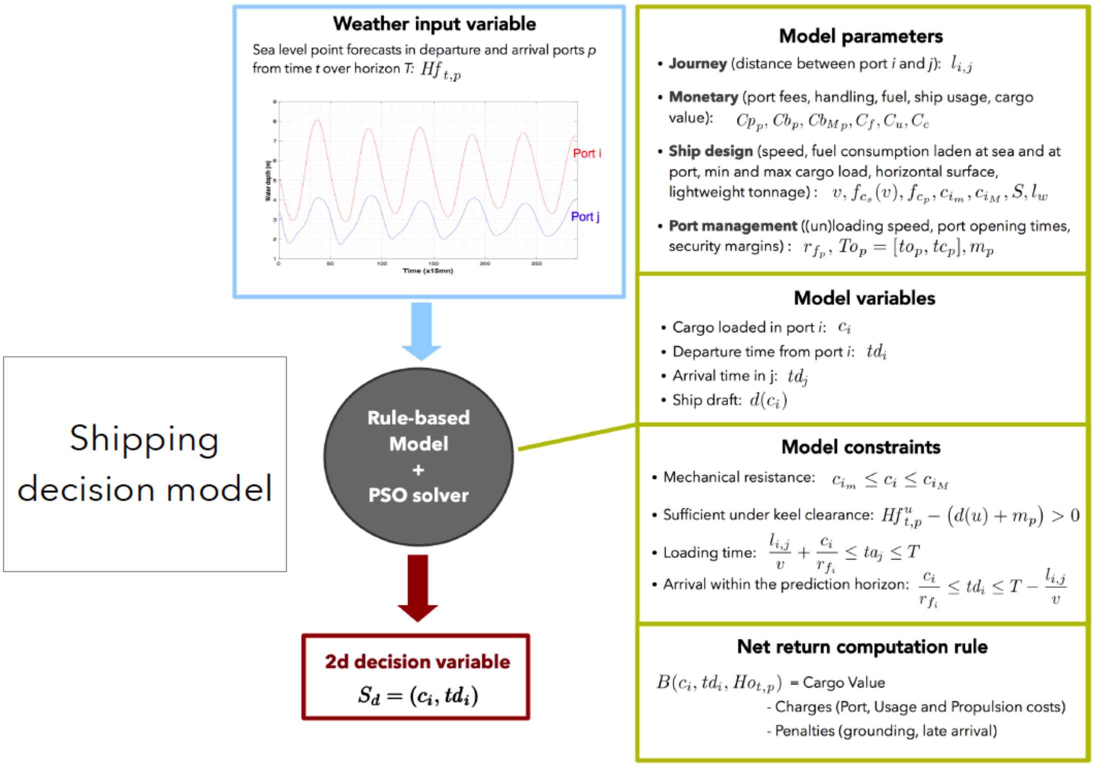

Weather input variable
Sea level point forecasts in departure and arrival ports *p* from time *t* over horizon *T*: $Hf_{t,p}$
Water depth (m)
1
2
3
4
5
6
7
8
9
0
50
100
150
200
250
Time (x15mn)
Port i
Port j
Shipping decision model
Rule-based Model + PSO solver
2d decision variable
$S_d = (c_i, td_i)$
Model parameters
• **Journey** (distance between port *i* and *j*): $l_{i,j}$
• **Monetary** (port fees, handling, fuel, ship usage, cargo value): $Cp_p, Cb_p, Cb_{M\,p}, C_f, C_u, C_c$
• **Ship design** (speed, fuel consumption laden at sea and at port, min and max cargo load, horizontal surface, lightweight tonnage) : $v, f_{c_s}(v), f_{c_p}, c_{i_m}, c_{i_M}, S, l_w$
• **Port management** ((un)loading speed, port opening times, security margins) : $r_{f_p}, To_p = [to_p, tc_p], m_p$
Model variables
• Cargo loaded in port *i*: $c_i$
• Departure time from port *i*: $td_i$
• Arrival time in j: $td_j$
• Ship draft: $d(c_i)$
Model constraints
• Mechanical resistance: $c_{i_m} \leq c_i \leq c_{i_M}$
• Sufficient under keel clearance: $Hf^u_{t,p} - (d(u) + m_p) > 0$
• Loading time: $\frac{l_{i,j}}{v} + \frac{c_i}{r_{f_i}} \leq ta_j \leq T$
• Arrival within the prediction horizon: $\frac{c_i}{r_{f_i}} \leq td_i \leq T - \frac{l_{i,j}}{v}$
Net return computation rule
$B(c_i, td_i, Ho_{t,p})$ = Cargo Value
- Charges (Port, Usage and Propulsion costs)
- Penalties (grounding, late arrival)

**Figure 2.1** Overview of the shipping decision model in the case $N = 2$ ports. The methodology is similar for $N > 2$, as described in the text.

Group, 2017a,b]. The simple price band framework allows us to simulate a range of situations: night vs days, week days vs weekends, bank holidays. Finally, the safety margin coefficient $\alpha$ in terms of legally required under-keel clearance to use the confined navigation channel of port $p$ is set to $10\%$ of the laden ship draft as this is usual practice at limited speeds [NOREL, 2014]. The open sea version would require adding a $30\%$ margin to the dynamical draft.

### 2.2.3.2 Sea level input variables

The sea level point predictions in each port are harmonic tide forecasts available online through the British Oceanographic Data Center portal. The time step, $\Delta t = 15$ minutes, sets a minimum bound on the resolution of our departure time solution. In real port conditions, cargo ships cannot be expected to be exactly on time. Reducing this lower bound would consequently not be realistic. As the tide height can change quickly, and because we are dealing with additional centimetres of under keel clearance, it would also not be judicious to increase $\Delta t$ too much. Indeed, the sea level within 1 hour (or even 30 minutes) could change significantly with respect to the small variations we are interested in. Consequently, a time step of $15$ seems a suitable trade-off.

#### 2.2.3.3 Model variables

The ship draft, a key element in shipping planning and realisation, is a function of the cargo load as well as the fuel mass $f(t)$ in tanks at the time $t$ of interest. Considering Archimedes' principle and the equilibrium of forces in a gravitational field, the draft $r$ can be estimated from the equality between ship's weight and water displacement. In a simple approximation (barge ship), this leads to:

$$r(t) = \frac{m + f(t) - 0.5m_{max}}{\rho S} + r_{50} \qquad \text{(Equation 1)}$$

where $r_{50}$ is the half laden ship's draft, $S$ the ship's horizontal area, $m_{max}$ its carrying capacity, $\rho$ the water density. The function $f(t)$ is computed by taking into account the fuel consumption rates at sea $f_s$ and at port $f_p$ respectively, the time already spent at sea and at port respectively at $t$, as well as the total fuel load necessary to move the ship from one port to another and (un)load material. Dynamical effects such as the squat effect or the heel due to the wind and the wave responses can reduce the under-keel clearance temporarily. They are not taken into account here beyond the safety margins $\alpha r(t)$ as, again, we consider the still water problem.

#### 2.2.3.4 Constraints

The ship's cargo and scheduling have to satisfy some constraints. First, at any stage, the cargo load $m_j$ cannot exceed the tank capacity $m_{max}$ and must fit with the requirements for safe structural behaviour of the hull ($m_j \geq m_{min}$), as well as with the demand constraints in the next ports to visit ($m_j \geq \sum_{k=j+1}^{N} a_k$). In the following: $m_{min}$ is taken as the minimum of the structural constraint and the economic constraint.

The fuel load necessary to carry the ship and its cargo $m_j$ over the distance $l = \sum_{k=j}^{N-1} l_k$ at speed $v$ and load/unload the freight at rate $u_p$ in port $p$ must be subtracted from $m_{max}$: $f_s l + f_p \sum_{p=j+1}^{N} T_p + m_j \leq m_{max}$, where the minimal time spent at port $p$ is the time for (un)loading: $T_p = \frac{|m_{p-1} - m_p|}{u_p}$ (noting that we set $m_0 = 0$).

Second, to enter/leave port $p_j$ at time $t$, the water depth must be greater than the ship draft plus the safety margin:

$$\hat{X}_p(t) - (1 + \alpha) r(t) > 0. \qquad \text{(Equation 2)}$$

Third, the ship cannot leave port $p_j$ before the cargo is (un)loaded and must arrive before the horizon $T$ is reached, so:

$$t_{j-1} + \frac{l_{j-1}}{v} + \frac{|m_j - m_{j-1}|}{u_{p_j}} \leq t_j \leq T - \frac{\sum_{k=j}^{N-1} l_k}{v}. \qquad \text{(Equation 3)}$$

#### 2.2.3.5 Shipping return

The problem is to find the optimal combination of decisions $\boldsymbol{d}^* = (t_j^*, m_j^*)$, $j = \{1, ..., N-1\}$ that maximises the net benefit $B$, where:

$$B(\boldsymbol{d}; \hat{X}_j(t), j = \{1, ..., N\}) = \begin{cases} V - (O + P + U) & \text{if delivered on time,} \\ Z & \text{otherwise.} \end{cases} \tag{Equation 4}$$

The gross value $V$ is the merchant value of the cargo:

$$V = C_c.m_1 \tag{Equation 5}$$

with $C_c$ the unit value of the freight. From there, we subtract the operational costs of the journey, starting from $t_0$ (time of decision) with an empty ship and finishing at $t_a + \frac{m_N}{u_N}$ after unloading the material in port $p_N$ where $t_a$ is the arrival time in the last port of call. These charges encompass the propulsion costs:

$$O = C_f \left( f_s T_s + f_p \sum_p (T_p + T_{p*}) \right) \tag{Equation 6}$$

where $T_s$ is the total time spent at sea and and $T_p$, $T_{p*}$ the total times spent at port $p$ within and outside normal work hours respectively and $C_f$ is the fuel unit price. Operational charges also include usage costs:

$$U = C_u \left( T_s + \sum_p (T_p + T_{p*}) \right) \tag{Equation 7}$$

with $C_u$ the hourly usage cost (staff) of the ship. Finally, port costs have to be included:

$$P = \sum_p \left( \left\lceil \frac{T_p + T_{p*}}{24} \right\rceil C_p + T_p C_{bp} + T_{p*} C_{bp*} \right) \tag{Equation 8}$$

where $\lceil \cdot \rceil$ is a ceiling operator and $C_p$, $C_{bp*}$, $C_{bp*}$, the daily port fee, hourly manutention prices in normal hours and outside normal hours in port $p$ respectively.

$Z$ is the cost of not making the delivery in time (i.e within the horizon $T$). Depending on the aim of the user, $Z$ can also take into account the negative externalities on the environment and society of a grounding ($Z \to -\infty$) or simply the loss for the shipper ($Z = -V - (O + P + U)$).

## 2.3 A probabilistic approach to decision making

Using the model described above, one can choose an optimisation technique (e.g. particle swarm optimisation or simulated annealing) to compute the optimal decision to take at time $t_0$, according to the sea level forecast time series $\hat{X}_j(t)$, for the ports $p_j$, $j = \{1, ..., N\}$. Such a calculation does not consider the actual stochastic behaviour of the water depth. Local sea levels are influenced by a range of factors, including weather. A residual $e_j(t) = X_j(t) - \hat{X}_j(t)$ between the predictions and the observations can lead to either a regret ($e_j > 0$: the shipper could have loaded more or departed earlier) or a loss ($e_j < 0$: in order to adjust to the actual water level the journey is delayed, or a grounding can happen). In other words, the resulting solution is risky. It does not tolerate a negative deviation to prediction nor port delays. In order to take account of the uncertainty on the output of a given decision, we must introduce a risk measure.

Risk is a polysemous notion. This is reflected in the many works that have been published in order to identify and classify the variety of definitions (from Kaplan and Garrick [1981] to Aven [2012] and Goerlandt and Montewka [2015]). In the field of maritime transportation, an analysis of risk-related publications spanning over forty years (1974-2014) led by Goerlandt and Montewka [2015] shows that the majority of the works rely on four definitions:

(a) Risk is the expected value of the loss;

(b) Risk is a combination of scenarios, their probability and the extent of their consequences, represented as a triplet;

(c) Risk is the possibility of a loss; or

(d) Risk is the probability of an undesired event.

Although simplistic, definition (a) has the advantage of easing comparisons between two options as the information about the possible scenarios and their consequences is synthesized into a single number.

In the present work, 'loss' takes the meaning of the loss in profit due to the fact that decision $\boldsymbol{d} \in \mathcal{D}$ is taken at time $t_0$ based on imperfect forecasts $\hat{X}_j \in \mathcal{X}$ of the environment state $X_j \in \mathcal{X}$. Let $F_j$ be the cumulative distribution function over $X_j$, which is conditional on information on the prior values of $X_j$ and possible other information. Let $\hat{F}_j$ be a predictive distribution of $X_j$ (that is a distribution over $\hat{X}_j$) provided by the forecaster at $t_0$. Let $\hat{X}_j(t)$ be a point forecast time series of $X_j(t)$ over time $[t_0, t_0 + T]$, $B(.,.) : \mathcal{D} \times \mathcal{X} \to \Re$ the utility function (namely the net benefit of the journey based on decision $\boldsymbol{d}$) and $y(\cdot) : \mathcal{X} \to \mathcal{D}$ an optimal action function defined by:

$$y(\hat{F}_j) = \arg\max_{\boldsymbol{d} \in \mathcal{D}} \left( \mathbb{E}[B(\boldsymbol{d}, \hat{X}_j)]_{\hat{F}_j} \right) = \arg\max_{\boldsymbol{d} \in \mathcal{D}} \int_{\mathcal{X}} B(\boldsymbol{d}, \hat{X}_j) d\hat{F}_j \qquad \text{(Equation 9)}$$

The loss function $L(.,.) : \mathcal{D} \times [0,1] \to \Re$ is then defined by Granger and Machina [2006] as:

$$L\Big(y(\hat{F}_j), F_j\Big) = B\Big(y(X_j), X_j\Big) - B\Big(y(\hat{F}_j), X_j\Big) \qquad \text{(Equation 10)}$$

for all $\hat{X}_j, X_j \in \mathcal{X}$. In other words, the utility of the decision made under uncertainty $B\Big(y(\hat{F}_j), X_j\Big)$ is compared to the utility resulting from the decision made under perfect knowledge of the future $B\Big(y(X_j), X_j\Big)$.

The loss associated with a given decision can only be evaluated *a posteriori* as it requires the knowledge of the exact future states of the environment, that are not known at the time of the decision. Hence the recourse to the expected loss which only requests the actual knowledge on the possible values of these future states. We consequently define the risk $R$ of taking a shipping decision $\boldsymbol{d}$ as:

$$R(\boldsymbol{d}) = \mathbb{E}\left[L\Big(\boldsymbol{d}, X_j\Big)\right]_{F_j} \qquad \text{(Equation 11)}$$

Looking more closely at the definition of the loss which we aim to minimise (the expectation over the space of sea level residuals), one can notice that minimising $\mathbb{E}\left[L\Big(\boldsymbol{d}, X_j\Big)\right]_{F_j}$ is equivalent to finding the decision $\boldsymbol{d}^*$ that maximises the expected benefit $\bar{B}(\boldsymbol{d}) = \mathbb{E}\left[B\Big(\boldsymbol{d}, X_j\Big)\right]_{F_j}$.

The decision minimising $R$ would be, from a frequentist viewpoint, the one that, on average, over a large number of journeys, produces the maximal net benefit. The theoretical expectation addresses both the feasibility and the performance (high return) of the candidate solution, since the cost $-Z \to \infty$ of a grounding would prevent any solution with the least probability of grounding to be returned as optimal.

## 2.4 Implementation

The problem of deterministic shipping optimisation was defined in Section 2.2.3.5. It consists of finding the decision $\boldsymbol{d}^* = (t_j^*, m_j^*)$, $j = 1, ..., N-1$ maximising the net benefit of the shipping given sea level forecasts $\hat{X}_j$, $j = \{1, ..., N\}$. Similarly, the risk minimisation problem consists of finding the decision maximising the objective function (or risk function) defined in Section 2.3.

Both are constrained 2-dimensional optimisation tasks whose objective functions are not continuous nor differentiable. As a result, classical analytical optimisation techniques cannot be used. Hence the call to derivative-free algorithms such as particle swarm optimisation to estimate $\boldsymbol{d}^*$. A range of other computational methods could have been implemented as well. However, PSO was chosen because it generally demonstrates good convergence and execution speed properties in addition to its simplicity of implementation. A review and comparison of the derivative-free approaches is provided in Rios and Sahinidis [2013]. PSO is an iterative stochastic optimisation technique that imitates the natural swarm behaviour of a bird flock [Eberhart and Kennedy, 1995]. At each iteration, the elements (particles) of the flock explore the search space in a semi-random way and evaluate the fitness (value of the function to optimise) of their positions. They share the information so that their next move is influenced by both their own findings and the findings of the other members of the swarm. The algorithm stops when the desired number of iterations is reached and the position with optimum fitness is returned. Algorithm 1 describes the procedure and our implementation choices.

Because the risk function defined in Section 2.3 cannot be written in closed forms due to the definition of the net benefit $B$, it is natural to turn to Monte Carlo simulations to estimate them, within the PSO procedure. Algorithm 2 shows the general approach, now referred to as $R_{PSO}$. $B_{PSO}$ refers to the "deterministic" optimisation of the shipping benefit (by means of Algorithm 1), that is without taking into account any uncertainty on the sea level forecasts (although technically PSO is a stochastic technique). Hereafter we name nominal state the forecasted sea-level state.

#### 2.4.0.1 Sampling

Mathematically $R$ is, within a constant, the expectation of the economic output of a given decision when the sea levels in both departure and arrival ports vary around their nominal state (the predicted one). Such a definition implies that $R$ is model-dependent: its accuracy depends on the quality of the modelling of sea level residual distributions. In this section, we present the results of an analysis of these residuals in both departure and arrival ports of our first case study, namely Portsmouth and Liverpool.

The dataset used for the modelling and then the testing consists of sea level residuals sampled every 15 minutes between 00:15-01/01/2006 and 23:45-31/12/2016 UTC, in each port. We split it into two parts: even years (dataset $D_e$) and uneven years ($D_u$). The former is used for modelling the sea level residuals by means of best-fit distributions. It is then used for the shipping optimisation procedure *per se*. Finally, $D_u$ is used as validation set, to perform simulations and gather statistics on the performance of the optimisation outputs. The whole framework is summarized in Figure 2.2.

**Algorithm 1** Particle Swarm Optimisation procedure

1. Initialise randomly the position $\boldsymbol{d}_i$ of each particle $i$ in the search space $\mathcal{D}$ and set their initial velocity vector to **0**.

2. For each step $j$:

   (a) For each particle $i$:

      i. Compute the objective function $f(\boldsymbol{d}_i)$ (i.e. the net benefit $B(\boldsymbol{d}_i, \{X_j(t), j = \{1, ..., N\}\})$ resulting from the shipping decision $\boldsymbol{d}_i$ with actual sea levels $X_j(t)$, or the expected benefit $\bar{B}(\boldsymbol{d}_i)$). This is the fitness of position $\boldsymbol{d}_i$.

      ii. Update the personal best $\boldsymbol{b}_i(j)$ of each particle i.e its position with optimal fitness among the set of previous iterations. Similarly, identify and update the global best $\boldsymbol{g}(j)$, that is the best solution among the positions visited by the whole swarm so far.

      iii. Move each particle according to the following equation of motion:

      $$\boldsymbol{d}_i(j+1) = \boldsymbol{d}_i(j) + \boldsymbol{\nu}_i(j+1), \qquad \text{(Equation 12)}$$

      where the velocity is defined by:

      $$\begin{aligned}\boldsymbol{\nu}_i(j+1) = \omega(j+1)\boldsymbol{\nu}_i(j) + &\\ c_1 R_1\Big(\boldsymbol{b}_i(j) - \boldsymbol{x}_i(j)\Big) &+ c_2 R_2\Big(\boldsymbol{g}(j) - \boldsymbol{x}_i(j)\Big).\end{aligned} \qquad \text{(Equation 13)}$$

      The cognitive $c_1$ and social $c_2$ coefficients are set up so as to optimise the ratio between individual exploitation and social interaction while the linearly decreasing inertia weight $\omega(j)$ limits 'velocity explosion'. The diagonal matrices $R_1$ and $R_2$ introduce stochasticity in the walk of the particles.

3. Stop when the maximum number of steps is reached or when there is no change in the global optimum for a given number of steps. Return the position with optimal fitness.

**Algorithm 2** Procedure $R_{PSO}$

1. Initialise randomly the position $\boldsymbol{d}_i$ of each particle $i$ in the search space $\mathcal{D}$ and set their initial velocity vector to $\mathbf{0}$.

2. For each step $s$:

   (a) For each particle $i$:

      i. Sample time series of tide residuals $e_j(t)$ in ports $j = \{1, ..., N\}$ from a distribution model. In this first study, we assume the residuals to be independent from port to port at a given time, and between any two given times at a given port. The spatial independence is not a strong assumption as long as ports are not too close. On the other side, the time independence can be discussed as on short durations the residuals show correlation.

      ii. Compute the net benefit $B(\boldsymbol{d}_i, \{X_j(t), j = \{1, ..., N\}\}$ for the simulated sea level conditions. These are given by the nominal state modified by the tide residuals, namely at port $j$: $X_j(t) = \hat{X}_j(t) + e_j(t)$.

      iii. Repeat steps 2(a)i to 2(a)ii until the number $N_s$ of simulated environments requested to compute the empirical risk function is reached.

      iv. Estimate the latter from the $N_s$ outputs of step 2(a)iii.

   (b) Move particles according to the general PSO procedure described in Algorithm 1, in the search space, step by step. Here the objective function to be maximised are the risk functions computed in step 2(a)iv, e.g. the expected benefit or the expected loss.

   (c) Stop when the maximal number of steps is reached or when there is no change in the global optimum for a given number of steps and return the position with optimal fitness.

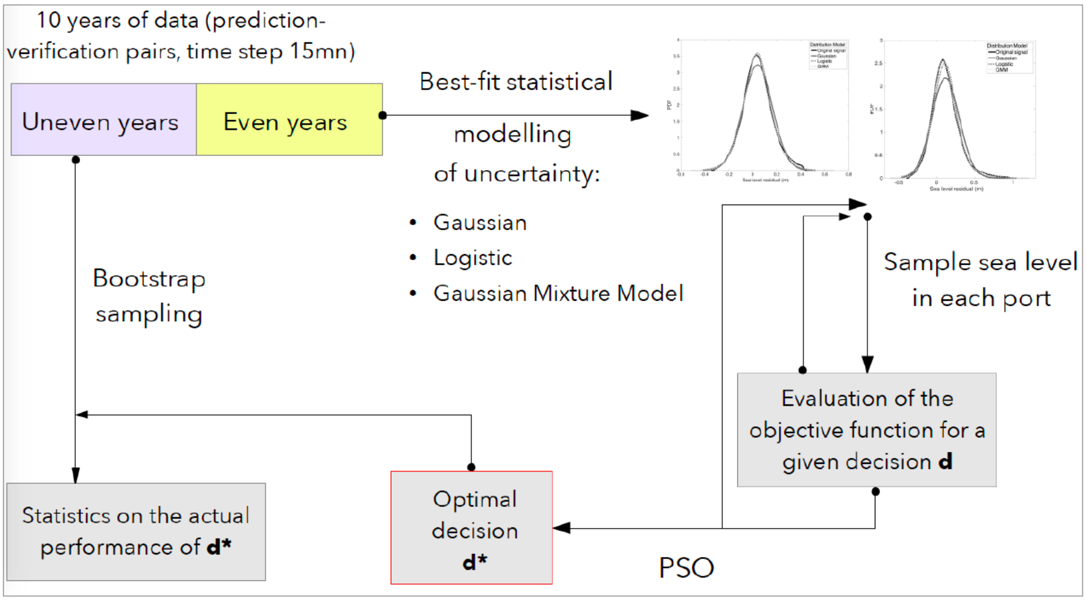
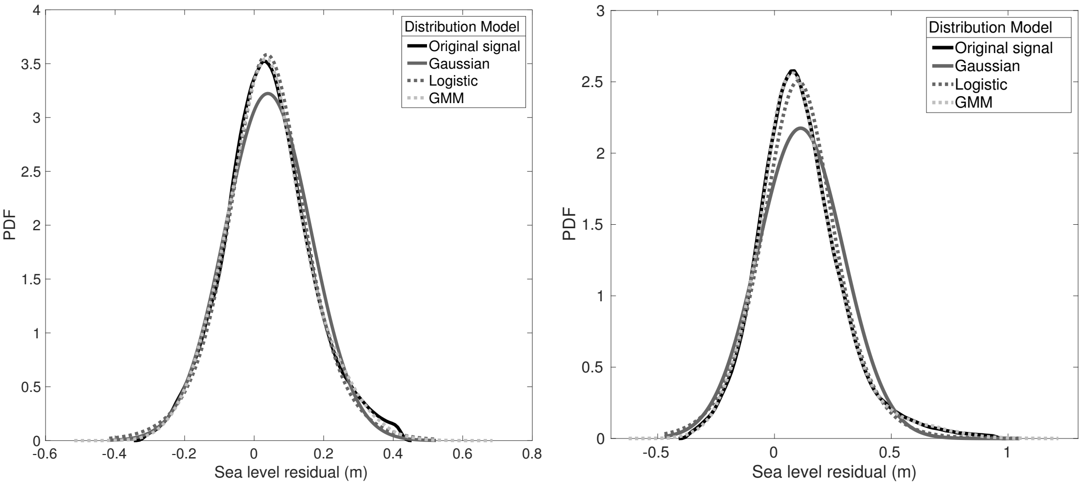

10 years of data (prediction-verification pairs, time step 15mn)
Uneven years
Even years
Best-fit statistical modelling of uncertainty:
• Gaussian
• Logistic
• Gaussian Mixture Model
Bootstrap sampling
Sample sea level in each port
Evaluation of the objective function for a given decision **d**
Statistics on the actual performance of **d***
Optimal decision **d***
PSO

**Figure 2.2** Methodological framework for data use in residual prediction and assessment of the resulting optimal decisions. The intertwining of the PSO algorithm with the Monte-Carlo sampling is also represented.

Three distributions were tested: Gaussian, Logistic and Gaussian mixture model (GMM). The number of components in the Gaussian mixture models were chosen so as to minimise the Akaike Information Criterion [McLachlan and Peel, 2004]. This criterion assesses the 'goodness of fit' of a model to a data set while introducing a penalty that increases with the number of free parameters requiring estimation. The aim is to find the optimal trade-off between model complexity and loss of information.

Kolmogorov-Smirnov statistics were computed to quantify the "goodness-of-fit" of each model. They reject at the 1% significance level the null hypothesis that the residuals follow a Gaussian or Logistic distribution for both ports. A graphical analysis of the three models shows that the Gaussian distribution significantly under-represents the small deviations of sea level observations with respect to tide predictions. Hence the introduction of the Gaussian mixture model, that globally represents the original residual distribution with greater fidelity. Besides, the GMM is able to capture the long tails that the single Gaussian or Logistic cannot. This could be important, as extreme events are usually in the tails.

This analysis will first be used to assess the distributional robustness of the optimisation procedures in Section 2.5, by analysing the effect of the residual modelling on the optimisation results. Besides, on a standard desktop computer running Linux, sampling from a Logistic distribution is about 10 times quicker than from a Gaussian distribution and 15 times quicker than from a 5 component mixture distribution. Since feasibility is at stake in operational research, in the following sections we check whether the difference in risk outputs and its implication in real-world decision making



**(a)** Portsmouth

**(b)** Liverpool

**Figure 2.3** Probability distribution functions of sea level residuals in Portsmouth and Liverpool ports over the period 01/01/2006-31/12/2016. Three models were fitted: Gaussian, Logistic and Gaussian mixture models.

justify the added complexity of the GMM input model.

## 2.5 Results and Discussion

Initially, we present the result of a study between $N = 2$ ports, allowing us to assess the distributional robustness of our probabilistic approach and justify implementation choices for the second study, a larger performance analysis with $N = 3$ ports.

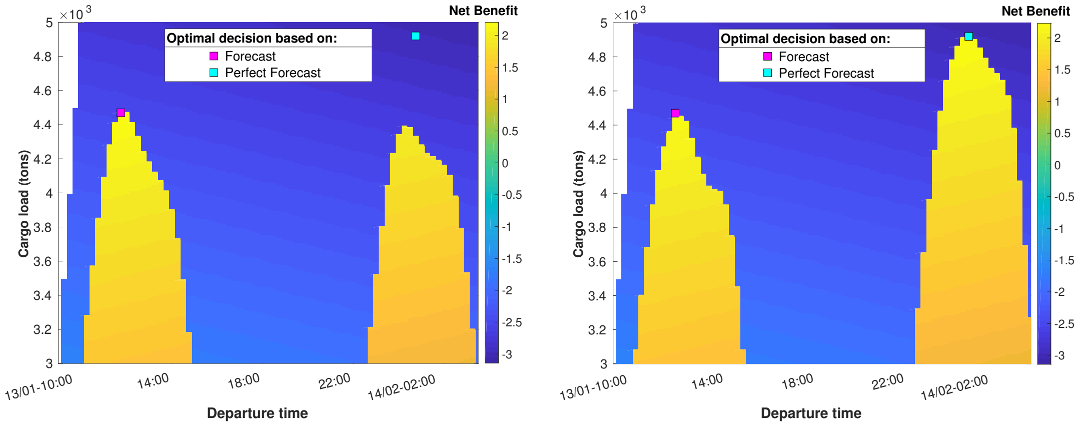
All the results in terms of benefit $B$ will be expressed as multiples of $B_0 =$ US\$ $363,550$ (resp. $B_0 =$ US\$ $190,530$ for the second case study). We also set the cost of not making the delivery in time to $Z = -V - (O + P + U)$. Negative benefits would thus imply a grounding or the impossibility to reach the arrival port within the specified time horizon.

### 2.5.1 Case-study 1: 2-port analysis

#### 2.5.1.1 Deterministic approach

The $B_{PSO}$ procedure recommends the ship to leave Portsmouth Harbour at $11:45$ UTC on January 13th 2017 with an overall barley freight of $4,475$ mt. The precision on these recommendations is estimated to $3$ mt in freight and $15$ minutes in time as standard deviations (from 1,000 independent runs).

Figure 2.4 presents a mapping of the final shipping benefit over the decision search space $\mathcal{D}$, given the forecast *a priori* at hand and given perfect forecasts, i.e. the *a posteriori* exact observations of the sea level depths. The optimal decisions according to $B_{PSO}$ in each scenario differ by one tide cycle in time and about $400$ mt in cargo



**(a)** Forecasted sea levels

**(b)** Actual sea levels

**Figure 2.4** Mapping of the net benefit $B$ over all the decisions $(t, m)$ of the search space, given sea level forecasts at hand (a) or actual sea level (b). The optimal decisions based on the deterministic forecasts and on the perfect forecasts (i.e. real state of the sea) through the solver $B_{PSO}$ are also reported.

load. In other words, the deterministic solution under imperfect harmonic predictions is far away from optimality in the real-world of non-zero residuals. Besides, it is quite straightforward to see on these maps that both solutions are very sensitive to perturbations. A 15mn departure/arrival shift or a negative error in the actual sea levels both shift the expected benefit from maximum to the negative area.

One way to get over the second limitation is to improve the accuracy of sea level forecasts. This is currently achieved by means of storm surge models. To take into account the local weather perturbations, these models use atmospheric forecasts as forcing in shallow-water hydrodynamic simulations e.g. the CS3 storm surge model covering the sea of the northwest European continental shelf [Flowerdew et al., 2010]. Nevertheless, whatever the accuracy reached, these forecasts cannot prevent the issue of port perturbations and delays. Hence it seems reasonable to develop a robust solution instead of a single deterministic optimisation.

### 2.5.1.2 Risk model

We now use $R_{PSO}$ to compute the optimal shipping decision under uncertain sea levels. The risk metric presented in Section 2.3 is combined with one of the three sea level residuals distribution models under consideration. Table 2.3 reports the statistical results of each combination as regards the optimal cargo load, departure time and the resulting guaranteed benefit at the error rate of 2 %, that is the 2% percentile $B_{.98}$. The latter is estimated from 100,000 Monte Carlo simulations. In order to prevent a methodological bias, these simulations sample the sea level by means of bootstrapping (over dataset $D_u$, c.f. Section 2.4.0.1).

As the purpose of the $R_{PSO}$ procedure is to support decision-making, it is necessary

**Table 2.3** Statistics over 50 runs of the outputs in terms of decision-making. The optimal cargo load $m$, departure time $t$ and guaranteed benefit $B_{.98}$ at the level of 2 % (over 100,000 simulations) are expressed in metric tons, UTC and fraction of $B_0$ respectively. The uncertainty is computed as the standard deviation of the results.

| **Distribution**<br>**Variable** | **GMM** | **Logistic** | **Gaussian** |
|---|---|---|---|
| $m$ (tons) | $4,210 \pm 20$ | $4,200 \pm 10$ | $4,225 \pm 9$ |
| $t$ (UTC) | $11:45 \pm 15$mn | $11:45 \pm 15$mn | $11:45 \pm 15$mn |
| $B_{.98}$ ($B_0$) | 2.069 | 2.064 | 2.077 |

to analyse the consequences of the above results as regards their translation in terms of practical shipping decision. The overall majority of the computed departure times are located within a 15 mn time slot centered on $00:15$. Taking into account the relative inertia of large vessels and generally slow port dynamics (from decision to subsequent actions), this range of uncertainty can be seen as a buffer to consider in the decision-making schedule. Trying to increase the precision on $t$ would be meaningless considering the real world context of a maritime shipping problem.

As regards the distribution impact, Logistic sampling produces more conservative loads than the GMM approach and further again, than the Gaussian one. The difference between the maximal and minimal loads above-mentioned is in the range of $25$ mt, that is in our case study less than 2 centimetres of draft. This leads to close guaranteed benefits $B_{.98}$.

Figure 2.5 summarises most of the information discussed above: a Logistic sampling will produce more stable (smaller variance) outcomes than the other residual models. It also shows that the approach can be said distributionally robust. Indeed, the ranges of the reduction in standard deviation and in guaranteed benefit when the underlying distribution varies are much smaller (close to 0.06 and 0.006 % respectively). Three observations can be highlighted as well. First, in this particular case study, the stochastic optimisation allows the owner to (in most of the configurations) save money as the guaranteed benefit is above the expected benefit of the deterministic decision in real conditions ($\bar{B} = 0.164$). Second, the spatial organisation of the points underlines a general pattern in robust optimisation: the guaranteed benefit increases at the cost of the increase in variance [Gotoh et al., 2015]. Finally, as noted by Gotoh et al. [2015], the variation in actual benefit is about one order of magnitude smaller than the reduction in its standard deviation.

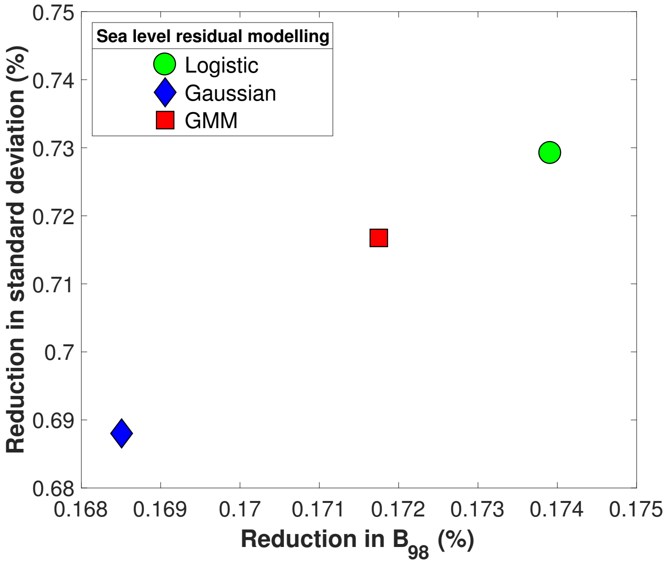


**Figure 2.5** Performance of each optimisation approach (sea level residuals distribution) from the perspective of the reduction of the guaranteed benefit at the error level of 2% and the standard deviation of the actual shipping benefit, with respect to the performances of the "deterministic" solution based on sea level forecasts alone. 100,000 Monte Carlo simulations are used to compute these statistics, with bootstrap sampling.

### 2.5.1.3 Summary of results

Figure 2.6 summarises all the above considerations in a 3-dimensional view of the optimisation problem. A map of the expected benefit (whose maximum corresponds to the minimal risk defined in the previous sections) is estimated with bootstrap sampling for each couple $(t, m)$ of the search space, as well as a map of the actual benefit variance on a smaller area of the search space. On top of both maps, are reported the decision suggested by the net benefit optimisation from sea level forecasts, perfect forecasts (i.e. perfect knowledge of the future) and our optimisation approach with a set of residual modellings.

Concretely, as the owner of the company, you could use the benefit optimisation decision that is based on the deterministic harmonic forecasts, load $4,475$ mt of barley and cast off at $11 : 45$. However the outcome of this decision, given the actual observations of sea levels is $-2.49B_0$. This is much less desirable than the benefit $2.12B_0$ that you could make if you knew the future perfectly and left Portsmouth port at 11:45 with $4,705$ mt on board. Using the stochastic optimisation method developed in this paper, you could load cargo between $4,200$ and $4,225$ mt, raise anchor at 11:45 and get a net benefit from $2.06B_0$ to $2.07B_0$.

If these decisions were reported in Figure 2.4(b) (mapping based on actual sea level conditions), one could notice that a port re-scheduling of up to 1.5 hours (earlier or delay) would not substantially change the benefit, nor a variation (in standard limits) in sea level conditions. Besides, Figure 2.6 reminds that the variance in the actual benefit

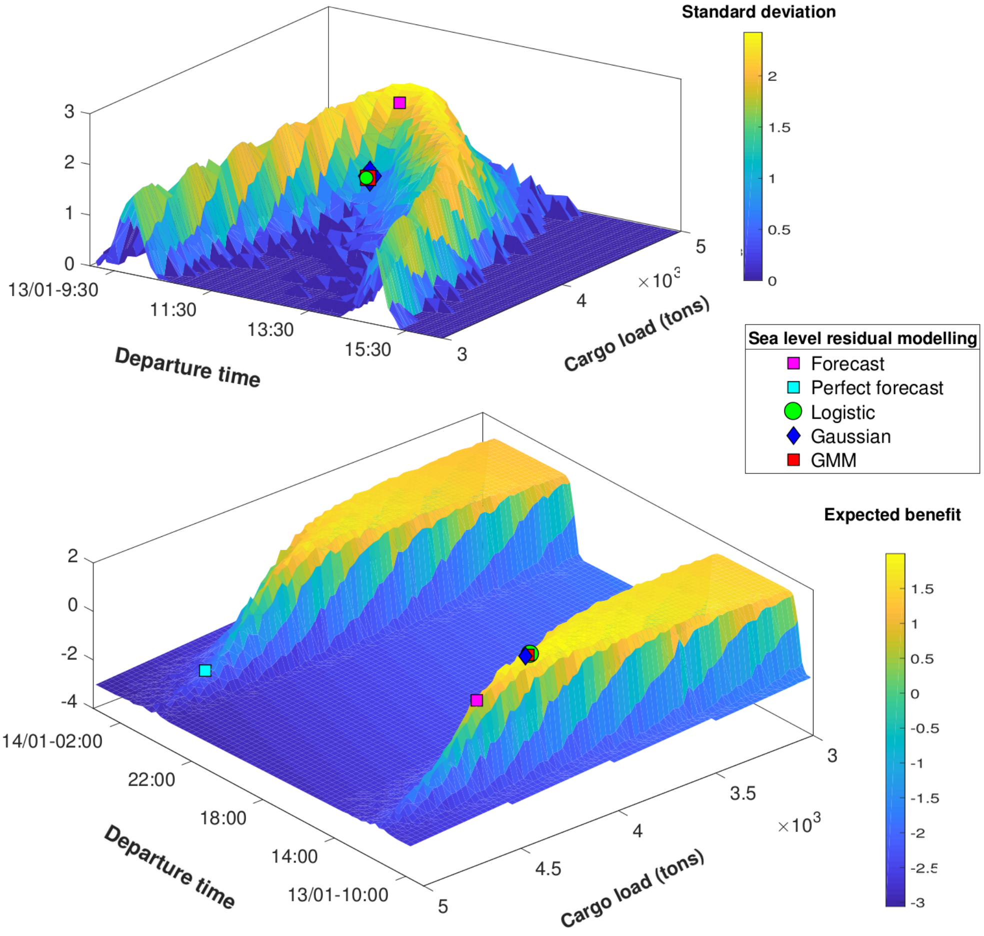
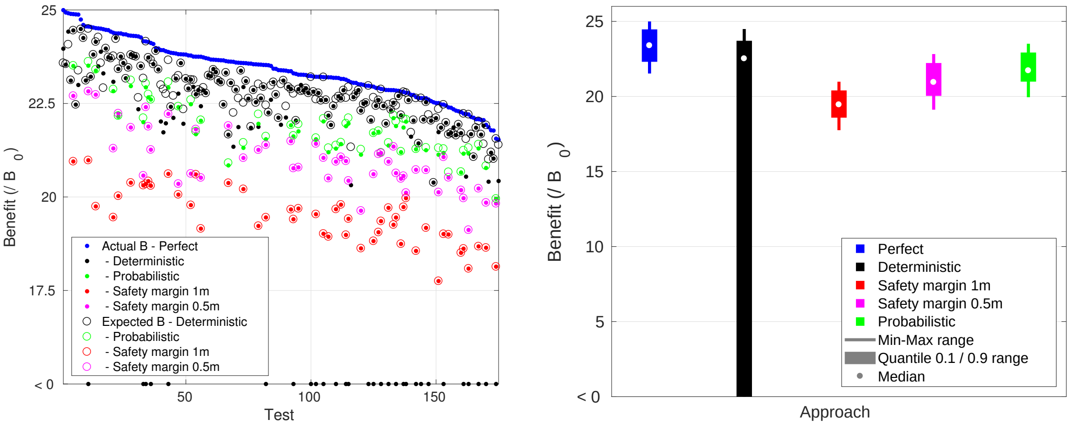


**Figure 2.6** Three dimensional mapping of each decision $(t, m)$ to the associated actual benefit standard deviation (top: zoom on the first high tide of the planning horizon) and expected net benefit (bottom, full planning horizon). Points of interest discussed in the text are also reported. The mapping use Monte Carlo simulations of 1,000 journeys by means of bootstrap re-sampling.



**(a)** **(b)**

**Figure 2.7** (a) Results over 175 decision times, $t_0$, and (b) a subset of 50 $t_0$ for the probabilistic and deterministic approaches, including the latter's 'rule-of-the-thumb' safe counterpart.

is substantially reduced for our solutions, contrary to the variance of the deterministic proposition. In other words, the approach $R_{PSO}$ proposes a robust solution. This is true for any sampling distribution although a Gaussian generally leads to solutions with slightly less predictable economic outcomes. Recalling the questions raised in the motivation of the problem (Section 2.1), in this case study, our stochastic approach demonstrated to be economically valuable with respect to the standard (deterministic) approach. Besides, a simple Logistic modelling of the residuals is enough to produce quality results, similar to those gained by means of a GMM.

One can note that the cargo load output $m$ can be turned into a safety margin $\Delta r$ to be deducted from the maximum draft that would have been allowed given the sea level tide forecasts at hand at $t_0$ (procedure $B_{PSO}$). For future works, it would be interesting to compare $\Delta r$ with what a "non-stochastic" commercial software would suggest on a similar problem, so as to assess the quality and potential added value of our model.

### 2.5.2 Case study 2: 3-port analysis

The first case study was a relatively simple example, chosen to show the potential of a probabilistic approach of tide routing, especially for tide-sensitive ports (Portsmouth in our illustration).

In the following, to provide a more representative analysis, the approach is applied to 175 different decision times $t_0$ between July 2016 and December 2016, each spaced by at least 24 hours. One additional port is also added to the analysis, with the chosen route: Liverpool-Portsmouth-Lowestoft. Again, we first compute the perfect decision, given a perfect knowledge of future sea level conditions in the three ports. The deterministic and probabilistic optimal decisions given tide predictions are then computed. Note that the probabilistic decisions are, given the results in the previous section, computed using logistic modelling of the sea level residuals. Besides, since this approach is more time consuming than a standard deterministic approach, we restrict the computations to 50 $t_0$, randomly sampled to represent the various trends in the whole set of 175 $t_0$. In addition to the deterministic decision, we add a deterministic decision taking into account a rule-of-the-thumb safety margin on the sea depth, as it is common practice in the maritime shipping industry. Static safety margins of 1m and 0.5m were both investigated in our experiments. We compare the performances of the three different approaches in terms of net benefit in actual conditions.

Figure 2.7(a) shows that 17% of the journeys cannot be fulfilled in the given horizon (i.e. cannot reach the final port, with the prescribed cargo load, because of low sea levels) if the deterministic decision is used without a safety margin. This score is lowered to zero by using the probabilistic approach or the deterministic one including a safety margin of 1 or 0.5m, confirming the robustness of our approach. Moreover, it is clear from Figure 2.7(b) that the net benefit in these 'critical' situations is significantly higher when using the probabilistic approach than the safe deterministic ones. This shows that switching from a traditional rule-of-the-thumb static safety margin, often too conservative, to a flexible safety margin provided by the probabilistic approach and taking into account the port calls to come, facilitates time savings and/or increased loading, improving the company's overall net benefit without foregoing safety. Such a solution can be said both robust and near-optimal.

### 2.5.3 Robustness to extreme sea level variations

One could argue that, as the probabilistic approach is based on the modeling of sea level residuals (itself fitted with archived observations), the results might be sensitive to extreme residuals. To analyse a possible lack of robustness to unseen sea level variations, we use the setting of case study 2. From the scheduling solutions for each of the 50 tests, we artificially modify the residuals in a time-window of $\pm 45$ minutes around the departure time in each port of transit. The perturbation procedure is the following: instead of the observed residuals, we sample them from the actual residual distribution (in each port of interest) whose mean is shifted to the 1) minimum residual ever observed; 2) quantile 0.1 of the residual distribution; 3) median of residuals; 4) quantile 0.9 of their distribution; 5) maximum residual ever observed. The impact on the net benefit of the journey is assessed and results are summarised in Figure 2.8. For the probabilistic approach, we measure the variation in benefit with respect to the unperturbed actual net benefit. For all deterministic approaches, the variation is measured with respect to the net benefit of the probabilistic approach resulting from the same perturbed conditions. Figure 2.8 shows that the net benefit resulting from the probabilistic approach is not sensitive to the more extreme residuals, whether negative

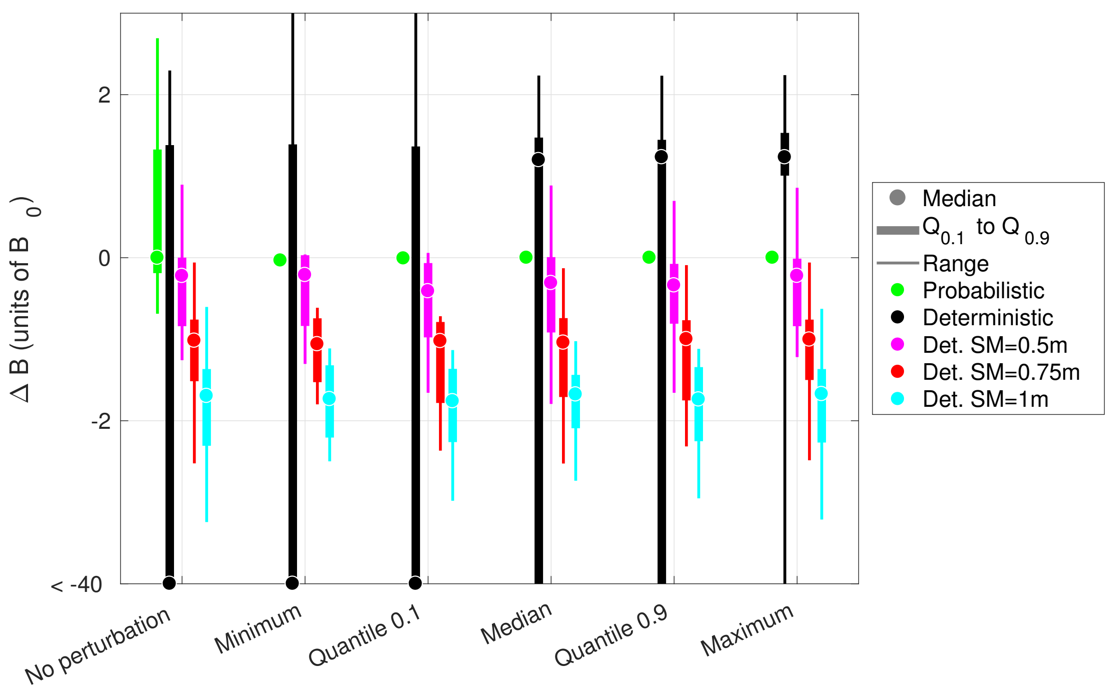


**Figure 2.8** Variation of the net benefit when the residual at each port of transit is artificially modified. We sample them from their original distribution whose mean is shifted to the value indicated along the x-axis. For the probabilistic approach (green), the variation in benefit is measured w.r.t. the unperturbed actual net benefit. For all deterministic approaches, the variation is measured w.r.t. the net benefit of the probabilistic approach resulting from the same perturbed conditions. $Q_{0.1}$ and $Q_{0.9}$ represent the quantile 0.1 and 0.9 of the set of results $\Delta B$ computed over the 50 tests.

or positive. The approach remains, in all conditions, more attractive regarding the actual benefit than the deterministic approaches with safety margins. Clearly, as a consequence of its conservative nature, the probabilistic approach cannot profit from the windfall effect generated by extreme positive residuals as much as a 0-safety margin approach.

### 2.5.4 Limitations

As stated in Section 2.3, the risk measure was chosen because its definition allowed us to address both feasibility and performance in terms of robust optimisation. However in practice, as detailed in the next section, $R(\boldsymbol{d})$ is estimated from Monte Carlo simulations of the shipping journey subject to various residual scenarios. These Monte Carlo simulations investigate a smaller uncertainty set than a theoretical expectation. The modelling of residuals is indeed based on historical data and potentially not conservative enough. Besides, calling Monte Carlo techniques implies that the number of sampled scenarios is limited, which is even more true if real-world applicability (computation time) of the decision-support tool is at stake.

We would like to conclude this section by further justifying one of the assumptions in our model. We chose not to consider the possible restrictions in terms of actual water depth during the loading or unloading steps. For more operational decision-making support, these additional constraints should be integrated. In our case study and generally speaking for small vessels, results are not affected by this simplification. As long as (un)loading rates are high and the loads small, the loading/unloading stages are very limited in time and the increasing ship draft matches the rising tide (which is the only potentially problematical scenario).

## 2.6 Conclusion and future works

This study introduced a decision model for robust cargo loading and ship scheduling in tidal areas. We associated a risk measure to each possible shipping decision. This measure was defined as the expected economic loss of taking the decision in an uncertain environment (sea levels), that is the loss with respect to the net benefit that could have been achieved if actual sea levels were perfectly known in advance. We developed a stochastic approach based on particle swarm optimisation and Monte Carlo simulations to estimate the decision that minimised risk. Results from a Portsmouth-Liverpool case study showed that this solution was both robust and optimal with regard to real port and sea level conditions. We also addressed the question of residual modelling and the resultant issue of distributionally robust optimisation. Thereon, the impact of a change of residual model on the optimisation outputs was negligible in terms of decision-making. A final application to 3 ports confirmed the added-value of our approach compared to standard practices.

While both the probabilistic and classical, ‘rule-of-thumb’, approaches can be considered robust (to, for example, port delays, forecasting errors etc.), the probabilistic approach was shown to be closer to optimality. Both case studies show the relevance of our approach for tide sensitive ports, small capacity carriers and cheap commodities. Finally, by analysing artificial extreme sea level variations, the robustness of this approach to unseen residuals and its efficiency over existing ’rule-of-the-thumb’ practices was demonstrated.

To address the computation time and underconservative historical modelling of residual issues, it would be interesting to define sounder uncertainty sets on which the risk metric would then be applied.

Another avenue of research is finer modelling of the sea level residuals, taking into account the cyclic character of data as well as results in the relevant literature, like those of Horsburgh and Wilson [2007] who noted patterns in the observation of highest weather-based surges. Because positive and negative deviations in tide prediction do not have the same effect on the end-user (that is the shipper), more attention could be given to their respective modelling as well as to the way of treating them through an appropriate asymetrical utility function, beyond what has already been done by focusing on the net benefit.

In practice, it will likely be necessary to make a more complex analysis. Shipping is a multi-dimensional activity. Loading / unloading a ship and leaving / entering a port require external support. We have analysed the robustness of shipping decisions under uncertain sea levels. However from congestion in waterways to berth availability, crane and tug allocation, a range of uncertain factors should also be included in the analysis of a robust optimal shipping decision.

Similarly, the uncertainty on the exact local water depth was assumed to come from the sea surface: the possibility of a weather-induced deviation to tides. Yet a range of factors can also locally modify in space and time the water depth: currents, sedimentation, vessel traffic for instance. Including the uncertainty on the lower part of the water column, at the sea floor, would consequently be interesting.

We assumed that the total fuel costs did not change significantly on a given journey when the cargo load is slightly increased. Our study would benefit from an analysis of the increase in fuel costs with the added cargo value as a function of the weather and ship characteristics (fuel consumption increasing with bad weather).

Finally, we intend to analyse our tide-routing problem from the perspective of existing weather routing solutions. The specificities of tide routing could be introduced in the dynamic criteria and constraints of such approaches.

# Bibliography

Mohammad Mahdi Abaei, Ehsan Arzaghi, Rouzbeh Abbassi, Vikram Garaniya, Mohammadreza Javanmardi, and Shuhong Chai. Dynamic reliability assessment of ship grounding using bayesian inference. *Ocean Engineering*, 159:47–55, 2018.

Agostinho Agra, Marielle Christiansen, Alexandrino Delgado, and Lars Magnus Hvattum. A maritime inventory routing problem with stochastic sailing and port times. *Computers and Operations Research*, 61:18–30, 2015.

Agriculture and Horticulture Development Board. Malting Barley UK Prices - Archives. http://cereals-data.ahdb.org.uk/archive/physical.asp, 2017. Accessed: 2018-01-07.

Terje Aven. *Foundations of risk analysis*. John Wiley and Sons, 2012.

Amir Azaron and Farhad Kianfar. Dynamic shortest path in stochastic dynamic networks: Ship routing problem. *European Journal of Operational Research*, 144(1): 138–156, 2003.

Volker Bertram. *Practical ship hydrodynamics*. Elsevier, 2012.

Michael J Briggs, Paul J Kopp, Vladimir K Ankudinov, and Andrew L Silver. Comparison of measured ship squat with numerical and empirical methods. *Journal of Ship Research*, 57(2):73–85, 2013.

Gerald G Brown, Robert F Dell, and R Kevin Wood. Optimization and persistence. *Interfaces*, 27(5):15–37, 1997.

Marielle Christiansen and Kjetil Fagerholt. Robust ship scheduling with multiple time windows. *Naval Research Logistics (NRL)*, 49(6):611–625, 2002.

Marielle Christiansen, Kjetil Fagerholt, Bjørn Nygreen, and David Ronen. Maritime transportation. *Handbooks in operations research and management science*, 14:189–284, 2007.

Tzung-Nan Chuang, Chia-Tzu Lin, Jung-Yuan Kung, and Ming-Da Lin. Planning the route of container ships: A fuzzy genetic approach. *Expert Systems with Applications*, 37(4):2948–2956, 2010.

Ali Dadashi, Maxim A Dulebenets, Mihalis M Golias, and Abdolreza Sheikholeslami. A novel continuous berth scheduling model at multiple marine container terminals with tidal considerations. *Maritime Business Review*, 2(2):142–157, 2017.

Hoda Davarzani, Behnam Fahimnia, Michael Bell, and Joseph Sarkis. Greening ports and maritime logistics: A review. *Transportation Research Part D: Transport and Environment*, 48:473–487, 2016.

Kinga Drwięga, Lucjan Gucma, and Rafał Gralak. Method for reserves determination of static and dynamic list of bulk carriers, applied to the dynamic under keel clearance system in the port of swinoujscie. *Annual of Navigation*, 24(1):89–102, 2017.

Yuquan Du, Qiushuang Chen, Jasmine Siu Lee Lam, Ya Xu, and Jin Xin Cao. Modeling the impacts of tides and the virtual arrival policy in berth allocation. *Transportation Science*, 49(4):939–956, 2015.

Russell Eberhart and James Kennedy. A new optimizer using particle swarm theory. In *Micro Machine and Human Science, 1995. MHS'95, Proceedings of the Sixth International Symposium on*, pages 39–43. IEEE, 1995.

Øyvind Endresen, Eirik Sørgård, Joachim Bakke, and Ivar SA Isaksen. Substantiation of a lower estimate for the bunker inventory: Comment on "Updated emissions from ocean shipping" by James J. Corbett and Horst W. Koehler. *Journal of Geophysical Research: Atmospheres*, 109(D23), 2004.

Jonathan Flowerdew, Kevin Horsburgh, Chris Wilson, and Ken Mylne. Development and evaluation of an ensemble forecasting system for coastal storm surges. *Quarterly Journal of the Royal Meteorological Society*, 136(651):1444–1456, 2010.

Wiesław Galor. Determination of dynamic under keel clearance of maneuvering ship. *Journal of KONBiN*, 8(1):53–60, 2008.

Floris Goerlandt and Jakub Montewka. Maritime transportation risk analysis: review and analysis in light of some foundational issues. *Reliability Engineering and System Safety*, 138:115–134, 2015.

Jun-ya Gotoh, Michael Jong Kim, and Andrew Lim. Robust empirical optimization is almost the same as mean-variance optimization. 2015. Accessed:2018-05-10.

Clive WJ Granger and Mark J Machina. Forecasting and decision theory. *Handbook of economic forecasting*, 1:81–98, 2006.

L Gucma and M Schoeneich. Probabilistic model of underkeel clearance in decision making process of port captain. *TransNav: International Journal on Marine Navigation and Safety of Sea Transportation*, 2(2), 2008.

Lucjan Gucma. Risk based decision model for maximal ship entry to the ports. In *Probabilistic Safety Assessment and Management*, pages 3022–3027. Springer, 2004.

KJ Horsburgh and C Wilson. Tide-surge interaction and its role in the distribution of surge residuals in the north sea. *Journal of Geophysical Research: Oceans*, 112(C8), 2007.

Stanley Kaplan and B John Garrick. On the quantitative definition of risk. *Risk analysis*, 1(1):11–27, 1981.

Elena Kelareva. The "DUKC Optimiser" ship scheduling system. In *2011 International Conference on Automated Planning and Scheduling System Demonstrations*, 2011.

Elena Kelareva. *Ship Scheduling with Time-Varying Draft Restrictions: A Case Study in Optimisation with Time-Varying Costs*. PhD thesis, The Australian National University, 2014.

Elena Kelareva, Sebastian Brand, Philip Kilby, Sylvie Thiebaux, and Mark Wallace. CP and MIP Methods for Ship Scheduling with Time-Varying Draft. In *Twenty-Second International Conference on Automated Planning and Scheduling*, ICAPS '12, 2012.

Eduardo Lalla-Ruiz, Christopher Expósito-Izquierdo, Belén Melián-Batista, and J Marcos Moreno-Vega. A set-partitioning-based model for the berth allocation problem under time-dependent limitations. *European Journal of Operational Research*, 250(3): 1001–1012, 2016.

N Le Carrer, S Ferson, and P. L Green. Optimising cargo loading and ship scheduling subject to uncertain sea levels. 8th Workshop on Reliable Engineering Computing, 2018.

Noémie Le Carrer, Scott Ferson, and Peter L Green. Optimising cargo loading and ship scheduling in tidal areas. *European Journal of Operational Research*, 280(3):1082–1094, 2020.

Oleg Makarynskyy, D Makarynska, Michael Kuhn, and WE Featherstone. Predicting sea level variations with artificial neural networks at hillarys boat harbour, western australia. *Estuarine, Coastal and Shelf Science*, 61(2):351–360, 2004.

Ioannis Mallidis, Eleftherios Iakovou, Rommert Dekker, and Dimitrios Vlachos. The impact of slow steaming on the carriers' and shippers' costs: The case of a global logistics network. *Transportation Research Part E: Logistics and Transportation Review*, 111:18–39, 2018.

Geoffrey McLachlan and David Peel. *Finite mixture models*. John Wiley and Sons, 2004.

NOREL. Under Keel Clearance - Policy Paper, 2014. UK Government.

Terry O'Brien et al. Experience using dynamic underkeel clearance systems: selected case studies and recent developments. In *30th PIANC-AIPCN Congress 2002*, page 1793. Institution of Engineers, 2002.

Peel Ports Group. Port charges: Port of Liverpool and Port of Manchester, 2017a.

Peel Ports Group. Schedule of Common User Charges: Liverpool Container Terminals, 2017b.

Quandl. Commodity prices. `https://www.quandl.com/collections/markets/commodities`, 2017. Accessed: 2018-01-07.

NM Quy, JK Vrijling, PH van Gelder, and R Groenveld. Parametric modeling of ship motion responses for risk-based optimization of entrance channel depths. In *11th World Conference on Transport ResearchWorld Conference on Transport Research Society*, 2007.

Luis Miguel Rios and Nikolaos V Sahinidis. Derivative-free optimization: a review of algorithms and comparison of software implementations. *Journal of Global Optimization*, 56(3):1247–1293, 2013.

Jin-Hwa Song and Kevin C Furman. A maritime inventory routing problem: Practical approach. *Computers and Operations Research*, 40(3):657–665, 2013.

Dongsheng Xu, Chung-Lun Li, and Joseph Y-T Leung. Berth allocation with time-dependent physical limitations on vessels. *European Journal of Operational Research*, 216(1):47–56, 2012.

Shucheng Yu, Shuaian Wang, and Lu Zhen. Quay crane scheduling problem with considering tidal impact and fuel consumption. *Flexible Services and Manufacturing Journal*, 29(3-4):345–368, 2017.

Lu Zhen, Zhe Liang, Dan Zhuge, Loo Hay Lee, and Ek Peng Chew. Daily berth planning in a tidal port with channel flow control. *Transportation Research Part B: Methodological*, 106:193–217, 2017.

## Chapter 3

# Possibly Extreme, Probably Not: Is possibility theory the route for risk-averse decision-making?

Given the simplicity of our initial sea level residual modelling and the existence of proper, physics-based predictions for the same residuals, we decided to investigate how to use the so-called ensemble predictions (EPSs) as input in our shipping optimisation framework. As described in depth in Section 1.1.2 of the introductive Chapter 1, we started to wonder about whether probability theory was the best way to extract information from the EPSs. A few preliminary works, based on actual residual data used in the shipping context, investigated non-probabilistic ways of modelling the residuals' EPSs.

Intervals bounded by the extremes of the EPS at hand in [Le Carrer, 2018a], as represented in Figure 3.1 were first used. The performance on a one-journey case study similar to the $N = 2$ -shipping optimisation problem developed in Chapter 2 were promising (cf. Figure 3.2) : the interval-based approach is less conservative than the probabilistic one and, in this setting, much quicker to compute. However, Figure 3.3 shows that if we run the same experiment over $64$ different starting days, we cannot generalise these results. When successful, the EPS-interval-based approach is the closer to optimality, however the failure rate is 34%, while the probabilistic approach developed in Chapter 2 is more conservative but has no failure, i.e. do not suggest a decision that will lead to unsufficient under-keel-clearance at some point of the journey (called failure due to its dramatic consequences). This is due to the overconfidence, and more generally lack of reliability, of the EPSs (i.e. their bound do not always contain the true value of the observed residual), largely documented in the literature [Buizza, 2018] (from model bias to too narrow EPS' variance). Although presenting the strongest potential w.r.t. predicting closest-to-optimal solution, such an interval modelling of the EPS cannot become operational. We consequently investigated in [Le Carrer, 2018b] (unpublished) the idea of combining to the EPS interval prediction a fuzzy safety margin derived from the probability, computed from logistic regression (with carefully chosen EPS or dynamical indicators), that a future residual will fall below the lower bound of the EPS.

Although performances (cf. Figure 3.4) were interesting, we did not go further

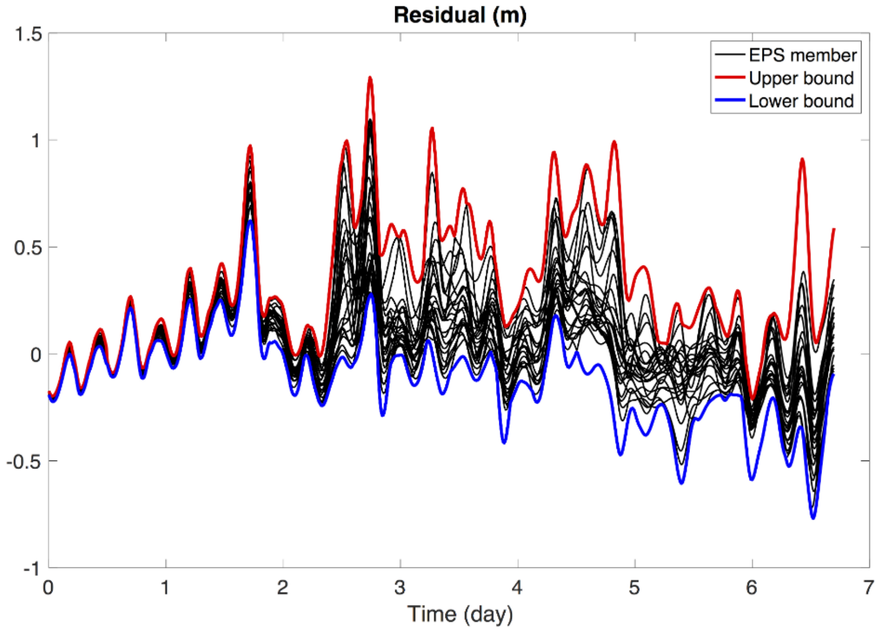


**Figure 3.1** Ensemble predictions (EPSs) for the residuals over the next 7 days, from July 1st 2017 6:00:00 GMT, at the port of Liverpool. An interval approach on residual prediction consists in saying that the future residual is comprised between the lower and upper bounds of the EPS at the time of interest.

in this direction as such a solution lacked of physical interpretability. It was more a correcting tool than a proper interpretation, which did not solve the problem raised by Bröcker and Smith [2008] and described in our Introduction (1.1.2.2). Rather, we decided to focus on the specific problem of EPS interpretation and considered a toy system regularly used in studies about ensemble forecasting strategies [Wilks, 2006, Williams et al., 2014], namely the Lorenz 96 system.

The possibilistic interpretation of EPS that we present in this PhD is first developed in the article *Possibly Extreme, Probably Not: Is possibility theory the route for risk-averse decision-making?* [Le Carrer, 2021], accepted in January 2021 in the journal *Atmospheric Science Letters*. It follows a first possibilistic tentative of interpreting EPSs, published in the proceedings of the 2019 Annual meeting of the European Meteorology Society [Le Carrer and Green, 2020]. We add in Appendix B extracts of this paper corresponding to the framework, a possibilistic dressing of ensemble members in place of the existing probabilistic ensemble dressing [Roulston and Smith, 2003]. Although results were interesting, we did not go further with this approach due to its parametric form, that implied a trade-off in performances as well as the impossibility to propagate the formal guarantees that possibility theory provides.

In this chapter, we investigate in particular the formal guarantees provided by our new framework, that we compare empirically with those provided by a classical probabilistic ensemble dressing of EPSs. The contribution of the authors is the following:

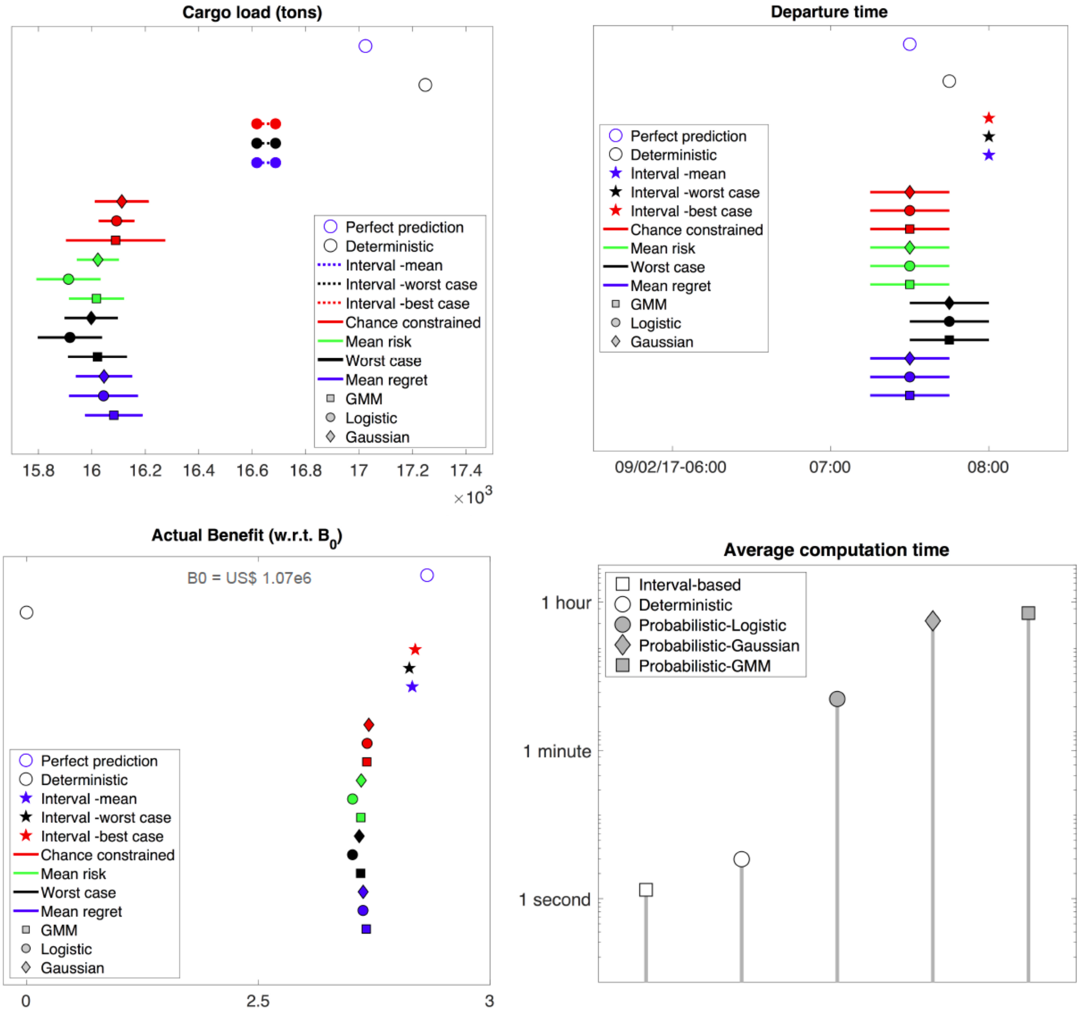


**Figure 3.2** Results from a case study with $N = 2$ (Liverpool-Southampton) ports, with setting (journey parameters) similar to the case study presented in Chapter 2, for which we try to optimise the best decision (cargo loading+departure time) to take on February 9th 2017 at 6:00:00 GMT when the sea level residual predictions are released by the MetOffice/NOC CS3 storm surge model.

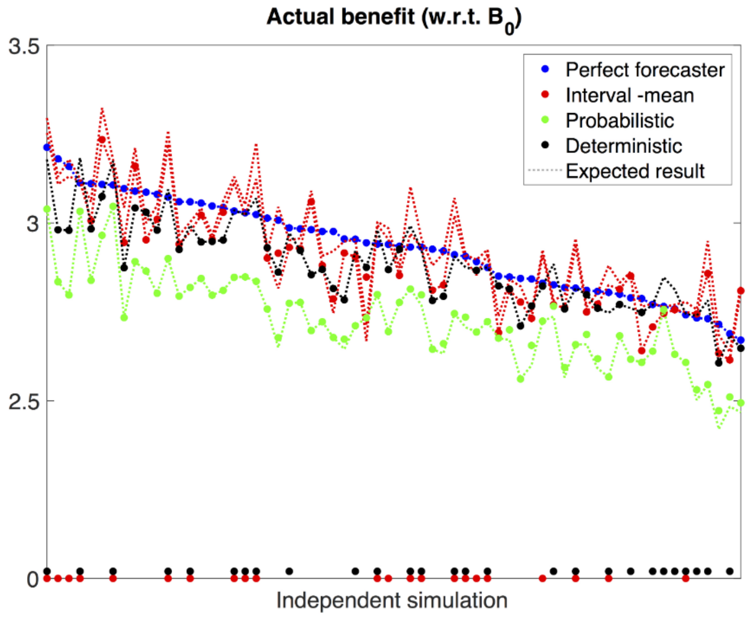


**Figure 3.3** Results of the experience described in Figure 3.2 over 64 successive days. We present the actual benefit depending on which optimisation method is used to take a scheduling decision. The expected benefits (resp. actual ones) are represented in dashed line (resp. solid points).

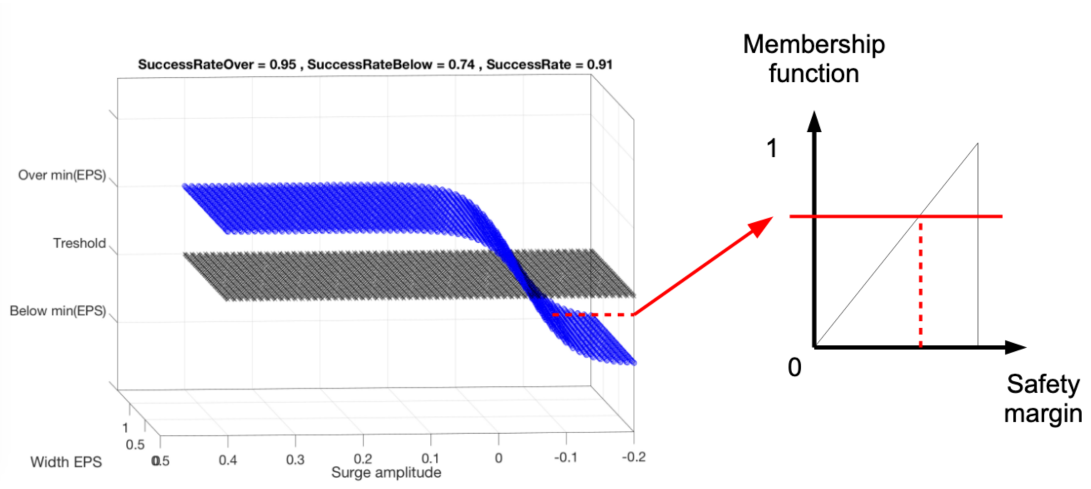


**Figure 3.4** Conceptual framework and performance of the logistic regression to predict out-of-EPS-bounds future sea level residuals from mean surge amplitude and EPS width and then turns the result into a fuzzy safety margin.

NLC developed the idea, designed the research, performed the experiments and wrote the article.

## Possibly Extreme, Probably Not: Is possibility theory the route for risk-averse decision-making?

N. Le Carrer

*Atmospheric Science Letters* (accepted January 2021)

**Abstract**

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

## 3.1 Introduction

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

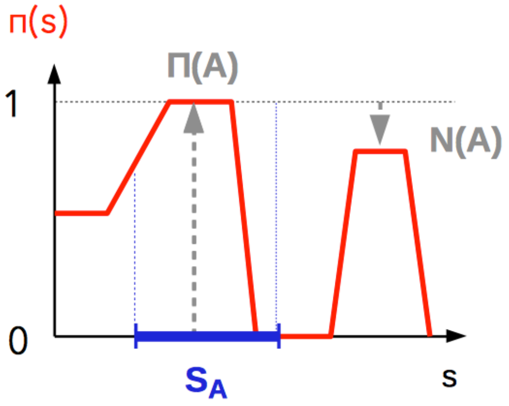


**Figure 3.5** Possibility distribution $\pi(s)$ where for an event of interest $A =$ “$s \in S_A$”, the possibility $\Pi(A)$ and necessity $N(A) = 1 - \Pi(\bar{A})$ measures are represented.

improving the prediction skills when it comes to common events, the actionability of probabilistic predictions often remains problematic [Smith, 2016]. In particular, the probabilistic prediction of extreme events often needs a development on its own [Friederichs and Hense, 2007, Friederichs et al., 2018].

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

Section 3.2 summarises the basics of possibility theory, Section 3.3 presents our possibilistic framework and discusses the theoretical guarantees that can be associated with its outputs. Section 3.4 introduces the synthetic experiments on the Lorenz 96 system [Lorenz, 1996] (L96) which allow us to assess these guarantees and their operational cost for both common and extreme events. We compare them with the outputs of a classical probabilistic interpretation of EPSs, and discuss our results in Section 3.5.

## 3.2 Possibility theory

Possibility theory is an uncertainty theory developed from fuzzy set theory by Zadeh [1978] and Dubois and Prade [2012]. It is designed to handle incomplete information and represent ignorance. Considering a system whose state is described by a variable $x \in \mathcal{X}$, the possibility distribution $\pi : \mathcal{X} \to [0, 1]$ represents the available information (or evidence) about the current state of the system. Given an event $A = \{x \in S_A\}$, where $S_A$ is a subset of $\mathcal{X}$, the possibility and necessity measures are defined respectively as: $\Pi(A) = \sup_{x \in S_A} \pi(x)$ and $N(A) = 1 - \Pi(\bar{A})$ where $\bar{A}$ represents the complementary event of $A$ (see Figure 3.5 for a visual understanding of these quantities). Both measures satisfy the following axioms and conventions [Cayrac et al., 1994]:

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

Possibility and probability distributions are interconnected through the concept of imprecise probabilities [Dempster, 2008]. A probability measure $P$ and possibility measure $\Pi$ are consistent iff [Dubois et al., 2004]:

$$P(A) \leq \Pi(A), \quad \forall\, A \qquad \text{(Equation 1)}$$

The definition of necessity implies that in these conditions:

$$N(A) \leq P(A) \leq \Pi(A), \quad \forall\, A \qquad \text{(Equation 2)}$$

.

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

As shown by Equation (Equation 2), a possibility distribution can be seen as a complete and consistent framework to deal with imprecise probabilities. Although the above procedure for computing a possibility distribution mostly relies on probabilities, its result contains more information than a purely probabilistic distribution *in the situation of incompleteness* (typically implied by a small dataset $S$). Indeed, the interval on the true probability allows the incompleteness of data or knowledge to be accounted for, while a point probability hides the fact that the said probability cannot be fully trusted (e.g. due to epistemic uncertainty). Figure 3.6 illustrates the results of this methodology applied to datasets sampled from a normal distribution, for various levels of $\beta$ and $N_s$. For a given $N_s$, the larger $\beta$ is, the more conservative is the distribution: $\gamma$ such as $\pi(x) \geq \gamma \ \forall x$ is larger, which implies that for any event $A \subset \mathcal{X}$: $\Pi(\bar{A}) \geq \gamma$. This also reads: $N(A) = 1 - \Pi(\bar{A}) \leq 1 - \gamma$, meaning that the confidence level associated with any $A$ cannot reach high values. Increasing $N_s$ reduces the relative effect of $\beta$ and all distributions tend in shape towards the underlying probability distribution, even if the tails remains more conservative for larger $\beta$.

## 3.3 Proposed framework

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

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

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

### 3.3.1 Deriving possibility distributions from EPSs

The objective of our possibilistic interpretation of EPSs is to derive from an EPS $\tilde{\boldsymbol{x}}_t$ and the archive $\mathcal{I}_t$ a possibility distribution $\pi(x_t|\tilde{\boldsymbol{x}}_t, \mathcal{I}_t)$, that encodes the knowledge

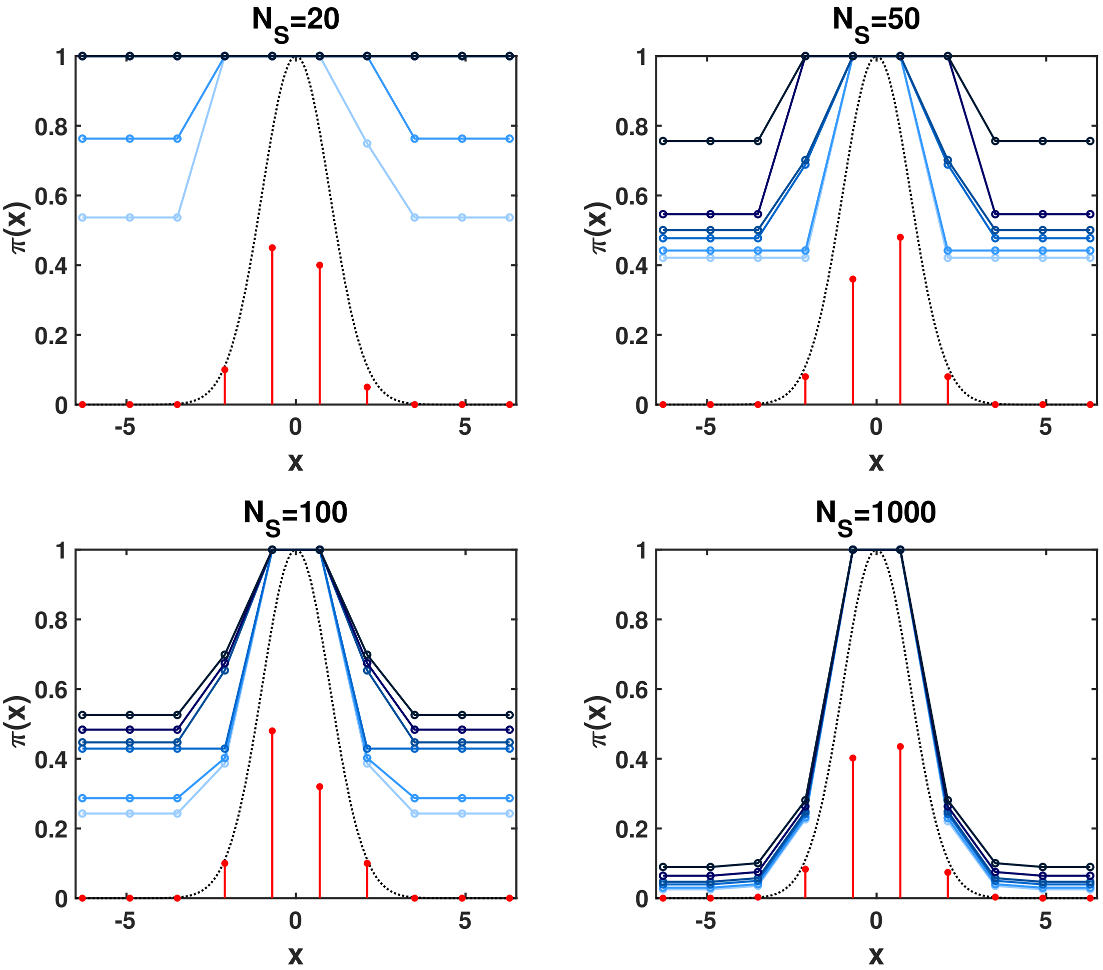
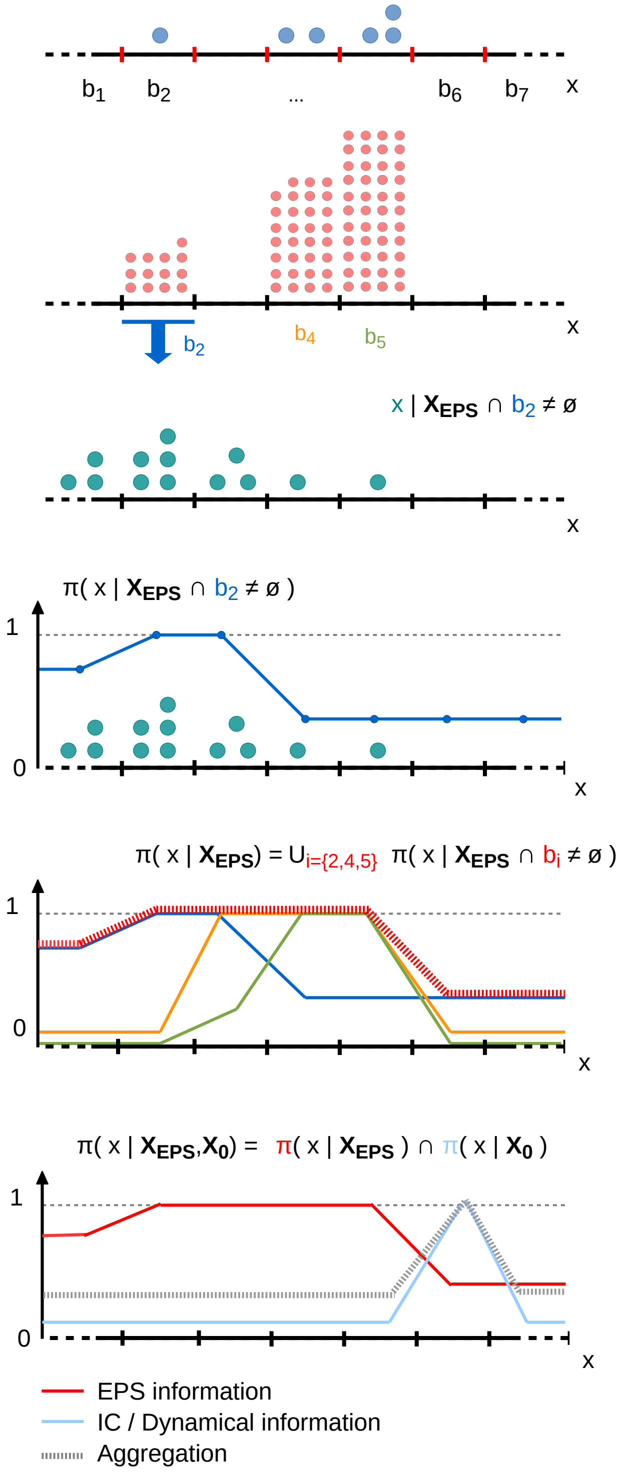


**Figure 3.6** Possibility distributions (solid lines) derived from datasets of $N_s$ elements sampled from a standard normal distribution. This derivation requires the computation of simultaneous confidence intervals for multinomial proportions over the $x$-axis binned into $n = 10$ bins. The effect of the confidence level $\beta = \{0.6, 0.75, 0.9, 0.95, 0.99, 1\}$ of the Goodman's formulation is shown (larger $\beta$ are plotted darker). Vertical red lines represent a frequency histogram of the same datasets and the normalised underlying Gaussian distribution is represented as a dotted line.



$b_1$
$b_2$
...
$b_6$
$b_7$
x
$b_2$
$b_4$
$b_5$
x
x | $\mathbf{X_{EPS}}$ ∩ $b_2$ ≠ ø
x
π( x | $\mathbf{X_{EPS}}$ ∩ $b_2$ ≠ ø )
1
0
x
π( x | $\mathbf{X_{EPS}}$) = $\cup_{i=\{2,4,5\}}$ π( x | $\mathbf{X_{EPS}}$ ∩ $b_i$ ≠ ø )
1
0
x
π( x | $\mathbf{X_{EPS}}$,$\mathbf{X_0}$) = π( x | $\mathbf{X_{EPS}}$ ) ∩ π( x | $\mathbf{X_0}$ )
1
0
x
EPS information
IC / Dynamical information
Aggregation

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 $I_t$ 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 3.7** Methodology of the possibilistic interpretation of EPSs developed in this paper.

derived from $\tilde{\boldsymbol{x}}_t$ about the verification $x_t$. The procedure described in this section is summarised and illustrated in the steps 1—5 of Figure 3.7.

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

To make use of the full set of ensemble members, we first partition the $x$-axis into $n$ bins $b_i$, take the subset $B$ of bins occupied by at least one ensemble member of the EPS, and compute $|B|$ possibility distributions $\pi(x_t|\tilde{x}_t^m \in b_j)$ where $b_j \in B$. Namely, for each bin $b_j \in B$ occupied by at least one ensemble member $\tilde{x}_t^m \in \tilde{\boldsymbol{x}}_t$, we retrieve the $N_s$ ensemble members $\tilde{x}_t^m \in b_j$ in the archive $\mathcal{I}_t$ and build a histogram of the set of corresponding verifications (so-called *analogs*) over the same binned $x$-axis. We then derive $\pi(x_t|\tilde{x}_t^m \in b_j)$ following the methodology presented in Section 3.2.

We obtain $|B|$ possibility distributions $\pi(x_t|\tilde{x}_t^m \in b_j)$, each dominating with confidence $\beta$ the true probability distribution $P(x_t|\tilde{x}_t^m \in b_j)$. Each possibility distribution provides the possibilities for the verification $x_t$ given the presence of one or more ensemble members in bin $b_j$ and is thus a partial view on the state $x_t$. Since there is only one truth for $x_t$ and several incomplete views on the verification, we can merge them through a disjunctive pooling [Dubois and Prade, 1992, Sentz et al., 2002]. Fuzzy set theory offers several definitions for computing the distribution resulting from the union of two fuzzy distributions. We adopt here the standard definition for its intuitive rationale: $\pi_{A\cup C}(x) = \max\big(\pi_A(x), \pi_C(x)\big)$.

We construct the resulting possibility distribution as:

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

### 3.3.2 From possibility distribution to prediction

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

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

$$P(x \in C_\pi^\alpha) = P(C_\pi^\alpha) \geq 1 - \alpha\ ,\ \forall\, \alpha \in [0,1]. \qquad \text{(Equation 4)}$$

This constitutes an easily verifiable consistency criterion [Hose and Hanss, 2019].

The possibility distribution satisfying this criterion is not unique. Beyond consistency, the choice of a possibility distribution to model the knowledge at hand is driven by the principle of maximum specificity [Dubois et al., 2004]. If $\pi_1$ and $\pi_2$ are two possibility distributions such that $\pi_1(x) \leq \pi_2(x)\ \ \forall x \in \mathcal{X}$, then $\pi_1$ is said more specific than $\pi_2$ and is more informative (i.e. less conservative). Maximum specificity w.r.t. the probabilistic information (*a priori* unknown) is achieved when the possibility distribution is probabilistically calibrated[1]:

$$P(C_\pi^\alpha) = 1 - \alpha\ ,\ \forall \alpha \in [0,1]. \qquad \text{(Equation 5)}$$

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

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

$$P\Big(P(x_t \in C_\pi^\alpha) \geq 1 - \alpha\Big) \geq \beta. \qquad \text{(Equation 6)}$$

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 $\beta$. Even for small sample sizes, the choice of $\beta$ is not critical and quasi perfect coverage rate is obtained: $\beta \geq 0.8$, ensures that $P\Big(P(x \in C_\pi^\alpha) \geq 1-\alpha\Big) \to 1$. Under this assumption, the $(1-\alpha)$-cuts can be used as candidate confidence intervals of *guaranteed* level $\alpha$. Ideally, we are looking for $(1-\alpha)$-cuts verifying Equation (Equation 5), which ensures optimal specificity of $\pi_{EPS}$ and thus maximally informative confidence intervals.

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

## 3.4 Experiments

### 3.4.1 Experimental setting

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

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

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

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

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

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

To reproduce the perturbation of the ICs, $M$ perturbed members $\tilde{X}_j$ are sampled independently around the true value of each variable $X_j$ following a normal distribution $\tilde{X}_j \sim \mathcal{N}(X_j, 0.1^2)$. These ensemble sets are initialised each time a new trajectory record starts, and integrated forward in time up to lead time $1.4$ by means of a Runge-Kutta $4$th-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 [MetOffice]).

### 3.4.2 Reference model: Gaussian ensemble dressing

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

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

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

Confidence intervals at level $\alpha$ on $x_t$ are obtained from $p$ by a method that provides the desired intervals associated with the highest-density regions [Hyndman, 1996]. We also report in the next section the performances of the confidence intervals similarly extracted from the unprocessed probability density (hereafter RAW) associated with the EPS (a histogram of the EPS normalised to represent a probability density).

### 3.4.3 Evaluation criteria

We aim at answering the questions:

(a) Can a possibilistic treatment of the EPS provide more guarantees than a probabilistic interpretation?

(b) If yes, at what cost?

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

### 3.4.4 Experiments

All results presented here use $n = 30$ bins of similar width to partition the $x-$axis[2]. The test set consists in $40{,}000$ independent trajectories of length $t = 7$ days and the corresponding EPS predictions. All EPSs have beforehand been preprocessed to remove the constant bias. We consider a range of archive size $N_I \in$

[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 $\approx 2$ 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.

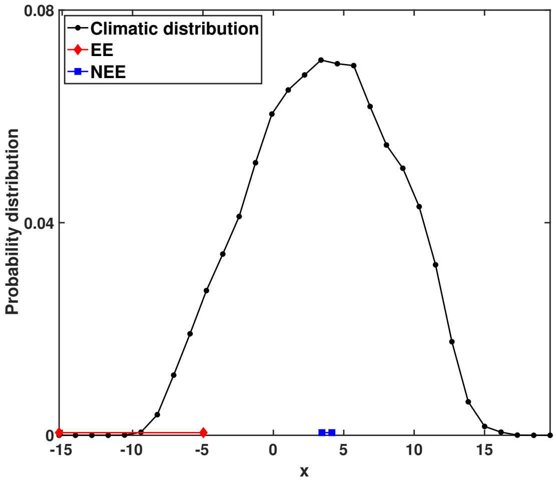


**Figure 3.8** Climatic distribution of the L96's variable of interest $X_1$ ($x$ for simplicity) where the 'extreme' event "$x \leq q_5$" (EE) and 'common' event "$q_{50} < x \leq q_{55}$" (NEE) are reported.

$\{156, 1560, 5 \times 10^3, 15 \times 10^3, 30 \times 10^3\}$. In particular, $N_I = 156$ corresponds to 3 years of model archive, whereas $N_I = 1560$ amounts to 30 years, which corresponds to the standard length of a historical re-forecast dataset [Hamill et al., 2004, Hagedorn, 2008]. The two latter $N_I$ are operational figures, unlike larger values that we present to study the asymptotic properties of our framework.

We define two types of events: an extreme event, "$x \leq q_5$" (EE), and a common event, "$q_{50} < x \leq q_{55}$" (NEE) where $q_i$ represents the percentile of level $i$ of the climatic distribution of $x$ (i.e. global distribution), plotted in Figure 3.8 along with both events. This will allow us to use test sets of similar sizes[3] in order to position our approach against the generic probabilistic postprocessing techniques that are known to weakly address such extreme events.

A preliminary assessment (Figure 3.9) of the effect of the parameter $\beta$ of Goodman's model on the probabilistic reliability of the $(1-\alpha)$-cuts derived from $\pi_{EPS}$ shows that varying $\beta$ from 0.6 to 1 does not impact guarantees at any given $N_I$ for the events of interest. It only impacts precision and its effect is only visible for small archives ($N_I \leq 156$) or large lead times, especially in the EE case. We consequently use $\beta = 0.9$ in our experiments, which allows to improve specificity while maintaining guarantees on confidence intervals.

## 3.5 Results

### 3.5.1 Empirical assessment of formal guarantees

Figure 3.10 reports the coverage probability of the confidence intervals $I^\alpha$ extracted for $\alpha \in \{0, 0.05, 0.1, \ldots, 1\}$ for all evaluated methodologies at lead times $t \in \{1, 3, 5, 7\}$

[3] About $2 \times 10^3$ elements.

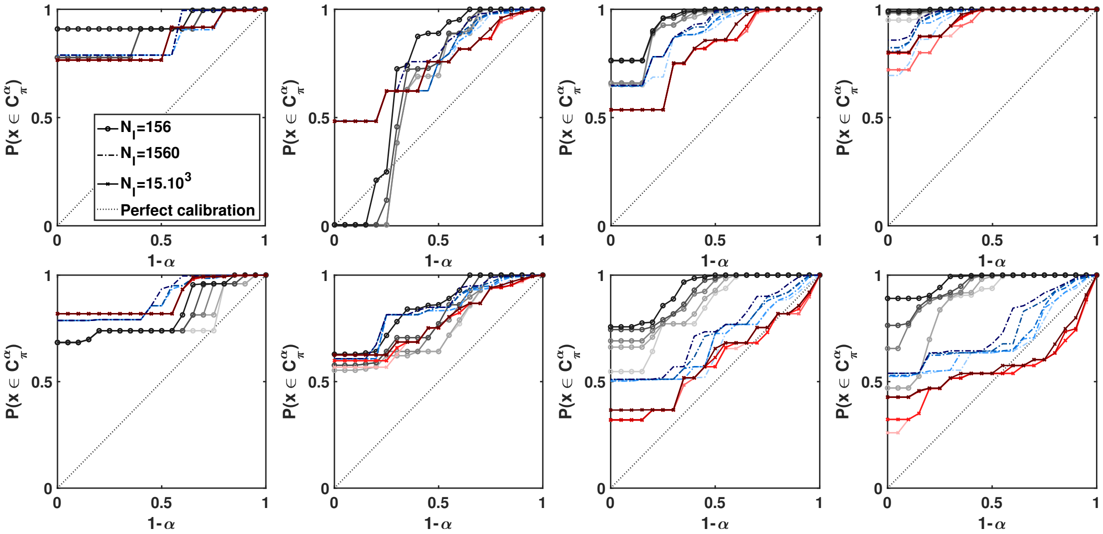


**Figure 3.9** Coverage probability of the $\alpha-$cuts of $\pi_{EPS}$ at lead time $t \in \{1, 3, 5, 7\}$ days (left to right), in the case of the NEE (top) and EE (bottom). Goodman models with parameter $\beta \in \{0.6, 0.75, 0.9, 0.95, 0.99\}$ (the darker the line, the larger $\beta$) are compared in the case of three archives of respective size $N_I \in \{156, 1560, 15 \times 10^3\}$ (grey, blue and red color scales respectively).

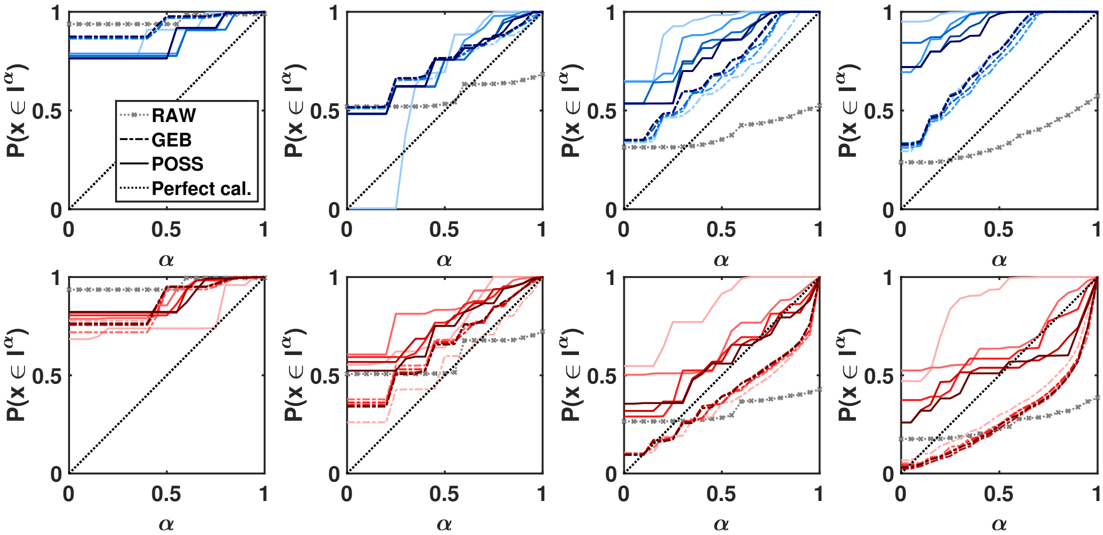


**Figure 3.10** Coverage probability of the $(1 - \alpha)$-cuts of $\pi_{EPS}$ used as confidence intervals of level $\alpha$ at lead time $t \in \{1, 3, 5, 7\}$ days (left to right), in the case of the NEE (top) and EE (bottom). The EPS archive size is $N_I \in \{156, 1560, 5 \times 10^3, 15 \times 10^3, 30 \times 10^3\}$ (the larger the darker the line). The coverage probability of the confidence intervals of level $\alpha$ derived from the raw EPS's probability density and from the postprocessed density (with the same training set of size $N_I$ as used in the possibilistic framework) is reported as well. The dotted diagonal represents perfect calibration.

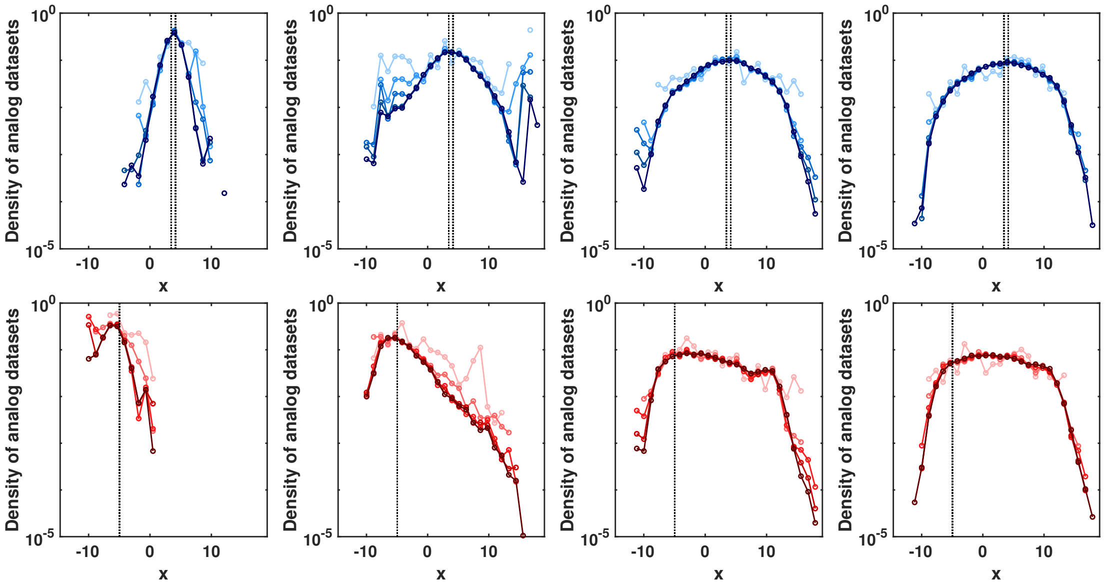


**Figure 3.11** Average density of the analog datasets used to derive $\pi_{EPS}$, for sizes $N_I \in \{156, 1560, 5 \times 10^3, 15 \times 10^3, 30 \times 10^3\}$ (the larger $N_I$ the darker the line) and lead time $t \in \{1, 3, 5, 7\}$ days (left to right), in the case of the NEE (top) and the EE (bottom). Only densities above $0$ are represented. Vertical dotted lines allow to visualise the events of interest (note that the EE is only defined by its upper bound).

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

We observe here a limitation of possibility theory: its strength lies in incomplete information. As shown in Figure 3.6, the larger the datasets used to derive possibility distributions, the closer the possibility distribution is in shape to the underlying probability distribution. In particular, the level $\gamma$ such as $\pi(x) \geq \gamma\ \forall x$ tends towards zero. In other words, such possibility distributions tend to conceal the possibility of rare events.

We illustrate this phenomenon in Figure 3.11, where we represent the average density of analogs used to compute the individual $\pi(x_t|\tilde{x}_t^m \in b_j)$ (see step 3, Figure 3.7). In the EE case, as the lead time increases, this average density decreases by several orders of magnitude for the more extreme bins ($x \to \inf \mathcal{X}$). This drop is all the more significant than $N_I$ is large. For small $N_I \leq 1560$, the more extreme bins are, as expected, not represented but the intermediary bins are and their density remains

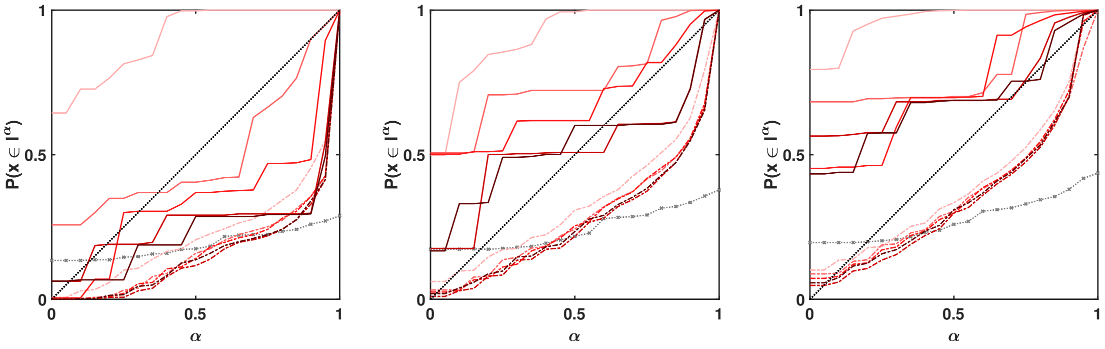


**Figure 3.12** Coverage probability of the $(1-\alpha)$-cuts of $\pi_{EPS}$ at lead time $t = 7$ days, in the case of events belonging to a partition of subsets of EE (from left to right: $x \leq q_1$, $q_1 < x \leq q_3$ and $q_3 < x \leq q_5$). The EPS archive size varies: $N_I \in \{156, 1560, 5 \times 10^3, 15 \times 10^3, 30 \times 10^3\}$ (the larger the darker the line). The probabilistic calibration of the confidence intervals of level $\alpha$ derived from the raw EPS's probability density and from the postprocessed density (with the same training set of size $N_I$ as the possibilistic framework) is also reported. See Figure 3.10 for legend.

above $\frac{1}{100}$. For very large $N_I \geq 5 \times 10^3$, the more extreme bins are represented however their density drops below $\frac{1}{1000}$. In other words, the rarest events part of EE are represented only for extremely large archives, where they will be part of large analog sets, which implies, given the asymptotic behaviour illustrated in Figure 3.6, that they will be concealed from the associated possibility distributions. More precisely, the level $\gamma$ such as $\pi_{EPS}(x) \geq \gamma \quad \forall x \in \mathcal{X}$ remains strictly positive so $P(x \in I^\alpha) = 1$ remains valid for $\alpha \approx 1$ (that is the large scale $(1-\alpha)$-cuts where $1-\alpha \rightarrow 0$). However for intermediate $\alpha$, the $(1-\alpha)$-cuts may not extend enough towards extreme bins, which negatively impacts the coverage rate. This trend is only observed for sufficiently large $\alpha$, as possibility distributions remain globally more conservative than the EPS-based probability distributions (see next Section), and consequently provide $I^\alpha$ that encompass more observations than the frequentist calibration requires in the case of smaller $\alpha$ (i.e. for the upper part of the distribution). The "sufficiently large $\alpha$" decreases with increasing lead times and archive sizes, following the effect described in Figure 3.11. Figure 3.12 illustrates our point by breaking down the coverage probability for three subsets of the EE: large archives lead to POSS-based confidence intervals that are all the more guaranteed as the event of interest is not too extreme. Probabilistic calibration for the more extreme part of EE can be improved by increasing the parameter $\beta$, however this has no effect in the case of large archives (see Figure 3.13).

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

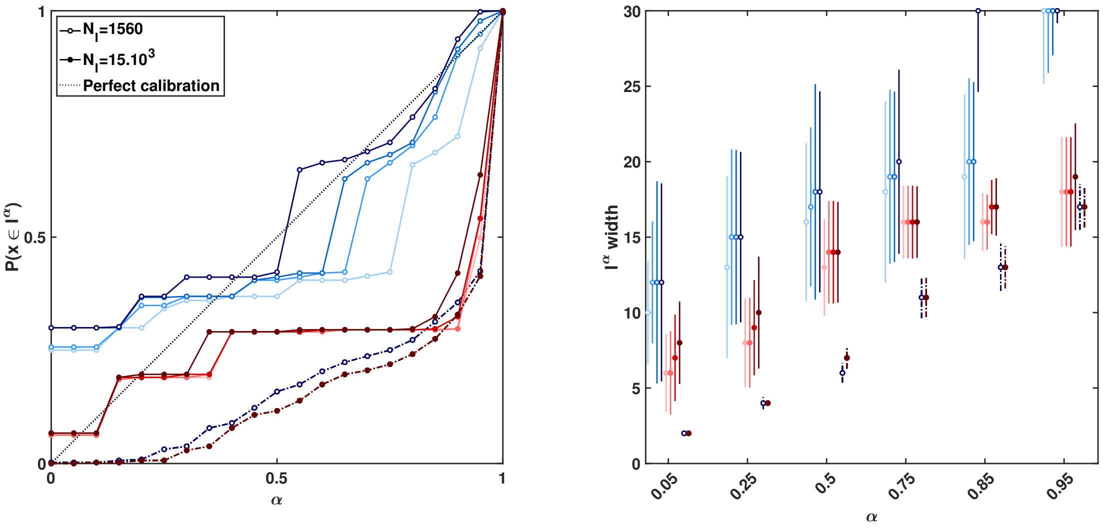


**Figure 3.13** Coverage probability (left) and associated width distributions (right) of the confidence intervals of level $\alpha$ at lead time $t = 7$ days, in the EE case, for two archive sizes $N_I \in \{1560, 15 \times 10^3\}$ (blue and red color scale respectively). POSS results (solid line) for increasing Goodman's parameter $\beta \in \{0.6, 0.9, 0.95, 0.99\}$ (the larger the darker the line) are compared to GEB results (dotted line). The width distribution is represented through its mean and one standard deviation above and below.

Provided that $N_I$ is not too large (which we assume is always the case for operational archives), Figures 3.10 and 3.12 clearly show the added value of treating the EPS in a possibilistic manner in terms of guarantees for the EE at large lead times, or in terms of reliability for the NEE at very small lead times. However, we can wonder what is the cost of such improvements. How do the possibility-based confidence intervals compare to their probability-based counterparts, in terms of precision?

### 3.5.2 Interval precision

Figure 3.14 compares the average width of the confidence intervals derived from the three methodologies. For both EE and NEE, $N_I$ affects the width of the possibilistic $I^\alpha$ significantly, making them narrower with larger $N_I$, all the more than the lead time increases. Their probabilistic counterparts are generally much smaller, except when $N_I \approx 30 \times 10^3$.

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

For EE, the added value of POSS over GEB is observed on two occasions: 1) intervals are as reliable yet narrower for very small lead times and $\alpha < 0.9$, whatever the archive

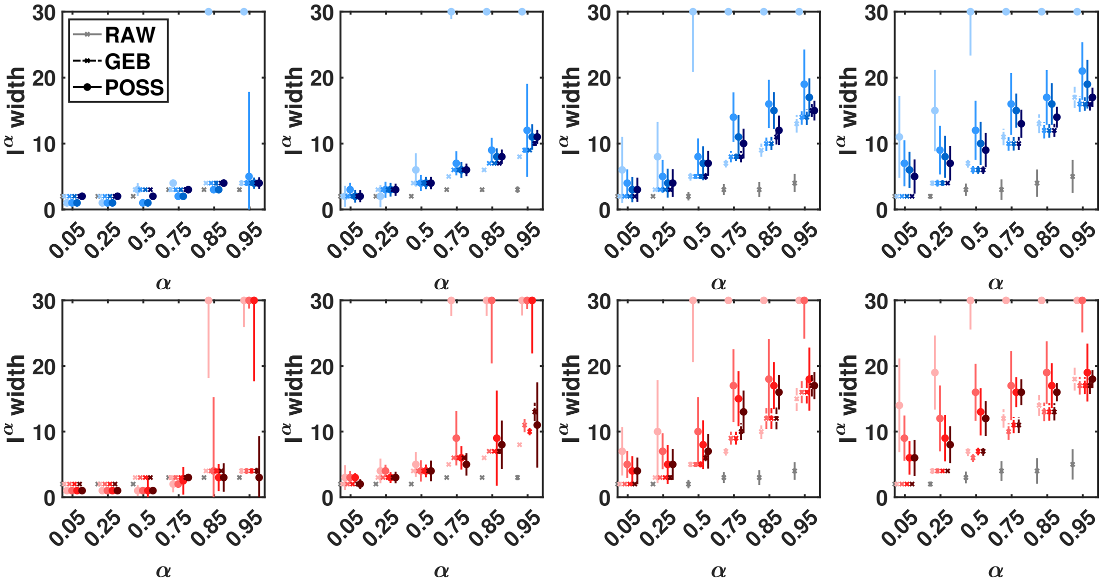


**Figure 3.14** Distribution (mean ± standard deviation) of the width of the possibility and probability-based confidence intervals described in the legend of Figure 3.10 for lead time $t \in \{1, 3, 5, 7\}$ days (left to right), in the case of the NEE (top) and EE (bottom). Only the cases $N_I \in \{156, 1560, 5 \times 10^3, 30 \times 10^3\}$ are represented (the larger $N_I$, the darker the line).

size ; 2) for large lead times and intermediary-sized archives ($N_I \in \{1560, 5 \times 10^3\}$), possibility-based confidence intervals are both guaranteed, reliable and operational (i.e. not too wide compared to GEB's results, contrary to what $N_I = 156$ produces), while the probabilistic intervals are narrower yet not guaranteed at all. In the case of particularly rare events, as represented in Figure 3.12, an intermediary archive such as $N_I = 1560$ is able to produce confidence intervals close to perfect reliability even for large lead times, as long as the parameter $\beta$ is increased towards 1. Such reliability is reached at the expense of the interval width, which is significantly increased (w.r.t. smaller $\beta$) for the largest $\alpha \geq 0.85$.

## 3.6 Conclusion

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

As raised by one of the reviewers of this study, in practice the verification (as observation) is a random variable itself [Tsyplakov, 2011, Lerch et al., 2017]. The use of confidence intervals rather than a Bayesian formalism and the derivation of credible intervals may consequently be discussed. Since our approach is taking such impreciseness into account (limited volume $S_{x_t}$ around $x_t$, Masson and Denoeux's transformation – cf. Section 3.3.1), even without explicitly tackling this problem, our framework accounts for (reasonable) randomness in the so-called verification.

Possibility theory is a promising tool for the prediction of extreme events, given a limited and imperfect amount of information on the system's dynamics. Beyond the results presented in this article, further developments by the author [Le Carrer and Ferson, 2020] show how $\pi_{EPS}$ can be combined with additional possibility distributions constructed from alternative sources of information such as the IC or dynamical information (see step $6$ of Figure 3.7). Therein, the concept of ignorance briefly introduced in Section 3.2 is developed and presented as an interesting tool for risk communication.

## Bibliography

Michael Scott Balch. New two-sided confidence intervals for binomial inference derived using walley's imprecise posterior likelihood as a test statistic. *International Journal of Approximate Reasoning*, 2020.

Jochen Bröcker and Leonard A. Smith. From ensemble forecasts to predictive distribution functions. *Tellus A: Dynamic Meteorology and Oceanography*, 60(4):663–678, 2008. doi: 10.1111/j.1600-0870.2007.00333.x.

Roberto Buizza. Ensemble Forecasting and the Need for Calibration. In *Statistical Postprocessing of Ensemble Forecasts*, pages 15–48. Elsevier, 2018. ISBN 978-0-12-812372-0. doi: 10.1016/B978-0-12-812372-0.00002-9.

D. Cayrac, D. Dubois, M. Haziza, and H. Prade. Possibility theory in "Fault mode effect analyses". A satellite fault diagnosis application. In *Proceedings of 1994 IEEE 3rd International Fuzzy Systems Conference*, pages 1176–1181. IEEE, 1994.

Arthur P. Dempster. Upper and Lower Probabilities Induced by a Multivalued Mapping. In Roland R. Yager and Liping Liu, editors, *Classic Works of the Dempster-Shafer Theory* *of Belief Functions*, pages 57–72. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008. ISBN 978-3-540-44792-4. doi: 10.1007/978-3-540-44792-4_3.

Didier Dubois and Henri Prade. On the combination of evidence in various mathematical frameworks. In *Reliability data collection and analysis*, pages 213–241. Springer, 1992.

Didier Dubois and Henri Prade. *Possibility theory: an approach to computerized processing of uncertainty*. Springer Science and Business Media, 2012.

Didier Dubois, Laurent Foulloy, Gilles Mauris, and Henri Prade. Probability-Possibility Transformations, Triangular Fuzzy Sets, and Probabilistic Inequalities. *Reliable computing*, 10(4):273–297, 2004. doi: 10.1023/B:REOM.0000032115.22510.b5.

P. Friederichs and A. Hense. Statistical Downscaling of Extreme Precipitation Events Using Censored Quantile Regression. *Monthly Weather Review*, 135(6):2365–2378, 2007. doi: 10.1175/MWR3403.1.

Petra Friederichs, Sabrina Wahl, and Sebastian Buschow. Postprocessing for Extreme Events. In Stéphane Vannitsem, Daniel S. Wilks, and Jakob W. Messner, editors, *Statistical Postprocessing of Ensemble Forecasts*, pages 127–154. Elsevier, 2018. ISBN 978-0-12-812372-0. doi: https://doi.org/10.1016/B978-0-12-812372-0.00005-4.

Tilmann Gneiting and Matthias Katzfuss. Probabilistic forecasting. *Annual Review of Statistics and Its Application*, 1:125–151, 2014.

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

R Hagedorn. Using the ecmwf reforecast dataset to calibrate eps forecasts. *ECMWF Newsletter*, 117:8–13, 2008.

Thomas M Hamill, Jeffrey S Whitaker, and Xue Wei. Ensemble reforecasting: Improving medium-range forecast skill using retrospective forecasts. *Monthly Weather Review*, 132(6):1434–1447, 2004.

Dominik Hose and Michael Hanss. Possibilistic calculus as a conservative counterpart to probabilistic calculus. *Mechanical Systems and Signal Processing*, 133:106290, 2019.

Rob J Hyndman. Computing and graphing highest density regions. *The American Statistician*, 50(2):120–126, 1996.

Noemie Le Carrer. Robust optimisation of cargo loading and ship scheduling in tidal areas. Annual Meeting of the European Meteorological Society, 2018a.

Noemie Le Carrer. Optimising ship scheduling subject to uncertain sea levels: Application to port traffic. https://d2k0ddhflgrk1i.cloudfront.net/CiTG/Over%20faculteit/Afdelingen/Transport%20%26%20Planning/Conferences/Matts/Le%20Carrer.pdf, 2018b. Conference Mathematics Applied in Transport and Traffic Systems.

Noémie Le Carrer. Possibly extreme, probably not: Is possibility theory the route for risk-averse decision-making? *Atmospheric Science Letters*, page e01030, 2021.

Noémie Le Carrer and Scott Ferson. Beyond probabilities: A possibilistic framework to interpret ensemble predictions and fuse imperfect sources of information, 2020. Under review.

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

T. P. Legg and K. R. Mylne. Early Warnings of Severe Weather from Ensemble Forecast Information. *Weather and Forecasting*, 19(5):891–906, 2004. doi: 10.1175/1520-0434(2004)019<0891:EWOSWF>2.0.CO;2.

Sebastian Lerch, Thordis L Thorarinsdottir, Francesco Ravazzolo, and Tilmann Gneiting. Forecaster's dilemma: Extreme events and forecast evaluation. *Statistical Science*, pages 106–127, 2017.

Edward N Lorenz. Predictability: A problem partly solved. In *Proc. Seminar on predictability*, volume 1, 1996.

Marie-Hélène Masson and Thierry Denœux. Inferring a possibility distribution from empirical data. *Fuzzy Sets and Systems*, 157(3):319–340, 2006. ISSN 0165-0114. doi: https://doi.org/10.1016/j.fss.2005.07.007.

MetOffice. The Met Office ensemble system. https://www.metoffice.gov.uk/research/weather/ensemble-forecasting/mogreps. Accessed: 2021-01-16.

K Mylne, C Woolcock, J Denholm-Price, and R Darvell. 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*, pages 113–118, 2002.

Mark S. Roulston and Leonard A. Smith. Evaluating Probabilistic Forecasts Using Information Theory. *Monthly Weather Review*, 130(6):1653–1660, 2002. doi: 10.1175/1520-0493(2002)130<1653:EPFUIT>2.0.CO;2.

Mark S. Roulston and Leonard A. Smith. Combining dynamical and statistical ensembles. *Tellus A: Dynamic Meteorology and Oceanography*, 55(1):16–30, 2003. doi: 10.3402/tellusa.v55i1.12082.

Kari Sentz, Scott Ferson, et al. *Combination of evidence in Dempster-Shafer theory*, volume 4015. Sandia National Laboratories Albuquerque, 2002.

Leonard A. Smith. Integrating Information, Misinformation and Desire: Improved Weather-Risk Management for the Energy Sector. In Philip J. Aston, Anthony J. Mulholland, and Katherine M.M. Tant, editors, *UK Success Stories in Industrial Mathematics*, pages 289–296. Springer International Publishing, Cham, 2016. ISBN 978-3-319-25454-8. doi: 10.1007/978-3-319-25454-8_37.

Alexander Tsyplakov. Evaluating density forecasts: a comment. *Available at SSRN 1907799*, 2011.

Daniel S Wilks. Comparison of ensemble-mos methods in the lorenz'96 setting. *Meteorological Applications*, 13(3):243–256, 2006.

R. M. Williams, C. A. T. Ferro, and F. Kwasniok. A comparison of ensemble post-processing methods for extreme events. *Quarterly Journal of the Royal Meteorological Society*, 140(680):1112–1120, 2014. doi: 10.1002/qj.2198.

L.A Zadeh. Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets and Systems*, 1 (1):3–28, 1978. ISSN 0165-0114. doi: https://doi.org/10.1016/0165-0114(78)90029-5.

## Chapter 4

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

The previous Chapter introduced a possibilistic framework to interpret EPS in a different way than the traditional probabilistic interpretations. It showed that when we are looking for reliable predictions (typically in the case of risk-averse decision-making), possibility theory offers guarantees, experimentally verified in the case of extreme events in particular, that a standard probabilistic interpretation do not. However, this has a cost when we deal with intermediate to large lead times (i.e. $\geq 3$ days), namely resolution. One solution is to take into account a so-far unused source of information, namely the initial conditions of the dynamical system (at the time where the predictive model is run). This is the point of this second paper on the possibilistic interpretation of EPSs: first, we show how to use a time series/monitoring of the dynamical (and to a large extent deterministic) system to derive a predictive possibility distribution on the future state of the system at a time of interest, by means of a similarity-based (or analog) method. Second, we show how combining this possibility distribution to the EPS-based possibility distribution by means of existing fuzzy rules, creates a synergy of information: from two conservative distributions, we manage to extract more information about the actual future state of the system.

Besides, contrary to Chapter 3 where we only focused on the continuous analysis of a possibility distribution, we now address the binary perspective (i.e. the case of binary predictions: $A$ *versus* $\bar{A}$). We thus investigate, when it comes to providing operational and user-friendly predictions, how to use at their full potential the possibilistic concepts of the dual necessity and possibility measures as well as the ignorance.

In this Chapter, we consequently report the article *Beyond probabilities: A possibilistic framework to interpret ensemble predictions and fuse imperfect sources of information*, submitted in July 2020 at the Quarterly Journal of the Royal Meteorological Society (minor revisions resubmitted in May 2021). The respective contributions of the authors are the following: NLD came with the research idea, designed the framework, the experiments and implemented them, analysed the results and wrote the article. SF reviewed the first version of the manuscript.

Before going further, we also add to this chapter the graphs summarising the performance in terms of coverage probability and width of the confidence intervals extracted from the possibility distribution fusing both EPS and dynamical information (COMB hereafter), according to the same experiments as presented in Chapter 3. This allows to assess the value of adding dynamical information, as well as the effect of the aggregation method (namely Zadeh's aggregation and the so-called general method, described in the article to come), on the continuous interpretation of possibility distributions presented in the previous Chapter 3.

On Figure 4.1, we can see that such an addition can significantly improve the reliability (by lowering conservatism) of predictive intervals for non extreme events (NEE) at intermediate and large lead times in the case of Zadeh's aggregation, all the more than the archive size $N_{I_A}$ is lower, yet to a certain extent ($N_{I_A} \approx 6$ months). The general aggregation method on the contrary make the intervals more conservative, again with the same dependence on the archive size.

When it comes to the extreme events (EE), we again observe that adding dynamical information to the EPS information by means of the general aggregation method does not lead to any improvement when it comes to the reliability of confidence intervals. The latter become generally more conservative. However, this is not true for larger lead times, where adding dynamical through such aggregation method for small $\alpha$, i.e large $(1 - \alpha)$-cuts (the peaks of the distribution) leads to an improvement in reliability of the associated confidence intervals. This is all the more true than the dynamical archive is small. On the contrary, Zadeh's aggregation of dynamical information allows to improve significantly the reliability of EPS-based confidence intervals for $t > 1$ day. However, the larger the lead time, the longer the dynamical archive needs to be (typically above $5$ years) to avoid unguaranteed confidence intervals at small $\alpha$.

What are the consequences in terms of confidence interval width? As shown on Figure 4.2, overall for both EE and NEE, adding dynamical information by means of Zadeh's aggregation tends to lower or maintain the interval's width while using the general aggregation will increase it. This is all the more true than the lead time is large and $\alpha$ small (i.e. that we are interested in the peaks of the distribution). This can be explained by the form of Zadeh's aggregation (min-envelope of two distributions), which implies that if one distribution is very conservative (e.g. EPS at large lead times) and the other (e.g. DYN) more peaked, as long as the latter peak matches with an area of high possibility for the former, only the information from the peaked distribution is kept in the aggregated distribution. That is how the interval width can decrease for small $\alpha$, at the expense of guarantees as noted on Figure 4.1.

Overall, we consequently advise to use in practice Zadeh's aggregation, which appears to be a reasonable trade-off between interval width (precision) and reliability (guarantees), in particular in the case of long dynamical archives (a few years). The general aggregation method tends to be over-conservative and do not facilitate synergy of information.

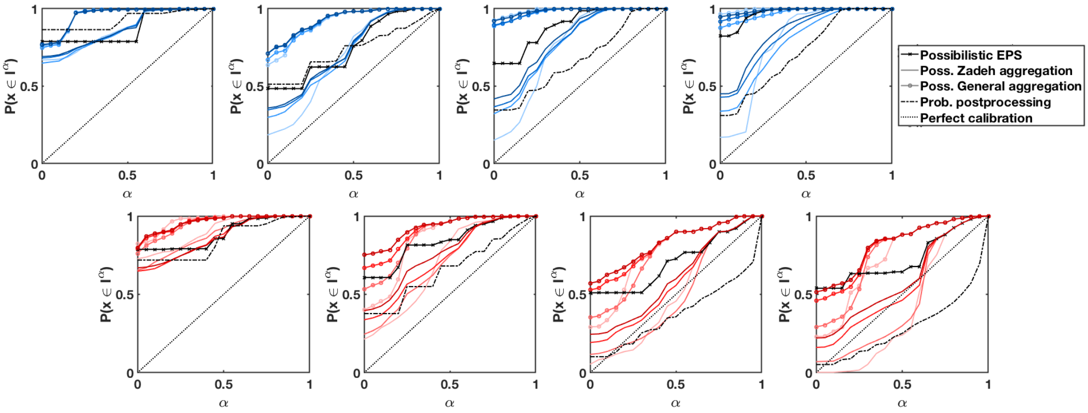


**Figure 4.1** Coverage probability of the $(1-\alpha)$-cuts of $\pi_{EPS}$ used as confidence intervals of level $\alpha$ at lead time $t \in \{1,3,5,7\}$ days (left to right), in the case of the NEE (top) and EE (bottom). The EPS archive size is $N_I = 1560$ while the time series' size, for dynamical information extraction, is $N_{I_A} \in \{2\times10^3, 2\times10^4, 2\times10^5, 2\times10^6\}$ (the larger the darker the line), corresponding respectively to system records of real-world equivalent duration of about 3 weeks, 6 months, 5.5 years 55 years. We compare the effect of two aggregation methods: Zadeh and the so-called general one. The coverage probability of the confidence intervals of level $\alpha$ derived from the postprocessed density (with the same training set of size $N_I$) is reported as well. The dotted diagonal represents perfect calibration.

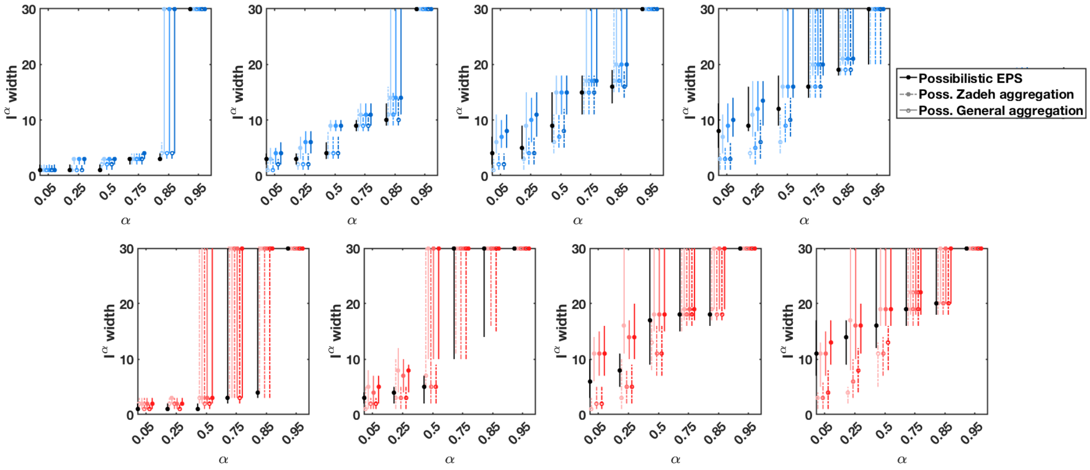


**Figure 4.2** Distribution (mean ± standard deviation) of the width of the possibility and probability-based confidence intervals for the approaches described in the legend of Figure 4.1 for lead time $t \in \{1,3,5,7\}$ days (left to right), in the case of the NEE (top) and EE (bottom).

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

N. Le Carrer and S. Ferson

*Quarterly Journal of the Royal Meteorological Society* (2nd revision submitted March 2021)

**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*

### 4.1 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 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 4.4. Section 4.2 introduces the basics of possibility theory, that we then use in Section 4.3 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 4.4. 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 4.5. A conclusion follows.

## 4.2 Possibility theory

### 4.2.1 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 $\pi$ is a function $\pi : \mathcal{X} \to [0, 1]$ that represents the

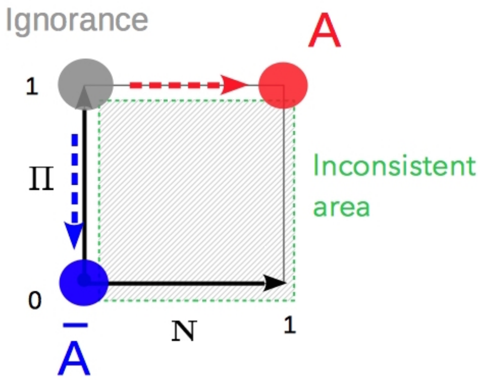


**Figure 4.3** $N - \Pi$ diagram, depicting the dual measures of possibility theory. $A$ is the event of interest and $\bar{A}$ its complement. The hatched area represents the area of inconsistent combinations for $N$ and $\Pi$.

state of knowledge about the current state 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 $A$. $\Pi$ 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 4.3. 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 $N(A) \rightarrow 1$, the more certain event $A$ is; the more $\Pi(A) \rightarrow 0$, 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 $\Pi$ 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:

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

Necessity and possibility measures can consequently be viewed as upper and lower bounds on the probability of a given event. Finally, we say that a possibility distribution $\pi$ is at least as specific as another $\pi'$ when $\pi(s) \leq \pi'(s) \quad \forall s \in \mathcal{X}$, in which case $\pi'$ is more conservative (or less informative) than $\pi$. 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].

### 4.2.2 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 $\beta$ 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 4.8). 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 $\beta$ 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 4.7 presents Masson and 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.

### 4.2.3 From possibility distribution to prediction

In this study, we focus on the binary interpretation of $\pi$, while the continuous interpretation is developed in Le Carrer [2021]. We are consequently interested in the prediction of an event $A$ of interest.

According to Section 4.2.1, we can extract from $\pi$ the possibility $\Pi(A)$ and necessity $N(A)$. Such measures provides coordinates to locate the corresponding point $\mathcal{P}$ in the $N - \Pi$ diagram sketched in Figure 4.3. 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 $\Pi$, $N$, or a function of both. However, using $\Pi$ 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. $\alpha$ 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 4.2.1). 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 $P_{IGN}$. 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-$\alpha$ the $P_\alpha$ 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). Other ways to turn the dual $\Pi$ and $N$ into a probability $P$ includes the pignistic transformation [Dubois et al., 2008]. However here we restrict the discussion to the options described above.

## 4.3 Framework

### 4.3.1 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 $x_{t_0+t}$.

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 $t_0$: $\tilde{\boldsymbol{x}}_{\boldsymbol{t_0+t}} = \{\tilde{x}^1_{t_0+t}, \dots, \tilde{x}^M_{t_0+t}\}$.

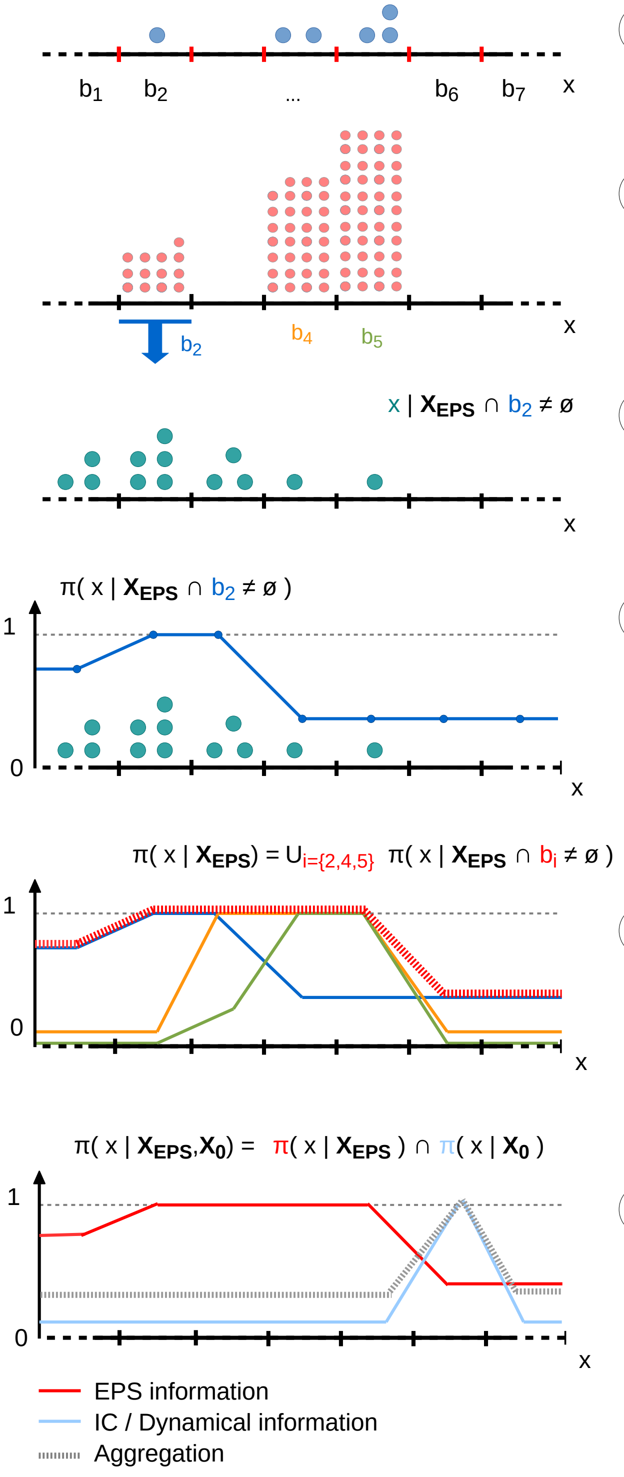
2. An archive $\mathcal{I}_t$ containing the pairs $\left(\tilde{\boldsymbol{x}}_{t_k+t}, x_{t_k+t}\right)$ for the lead time $t$ of interest and $N_I$ different starting time $t_k$ , $k = 1, \dots, N_I$. These instances are chosen so that the initial points $x_{t_k}$ and $x_{t_{k+1}}$ of two successive 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 $\approx 1$ 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.

### 4.3.2 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 4.4.

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


b1
b2
...
b6
b7
x
b2
b4
b5
x
x | XEPS ∩ b2 ≠ ø
x
π( x | XEPS ∩ b2 ≠ ø )
1
0
x
π( x | XEPS) = Ui={2,4,5} π( x | XEPS ∩ bi ≠ ø )
1
0
x
π( x | XEPS,X0) = π( x | XEPS ) ∩ π( x | X0 )
1
0
x
EPS information
IC / Dynamical information
Aggregation

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 $I_t$ 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 4.4** Step by step illustration of our framework.

$\pi(x_{t_0+t}|\tilde{x}^m_{t_0+t} \in S_x)$. 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 4.2.2, 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, $\{b_i, i = 1, \dots, n\}$.

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 $\beta$). 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:

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

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, $\pi_{EPS}$ 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.

### 4.3.3 Taking dynamical information into account

#### 4.3.3.1 Attractor reconstruction

The procedure of attractor reconstruction consists for a dynamical system characterised by a variable $x_t$ in finding the time delay $\tau$ and embedding dimension $m$ such that

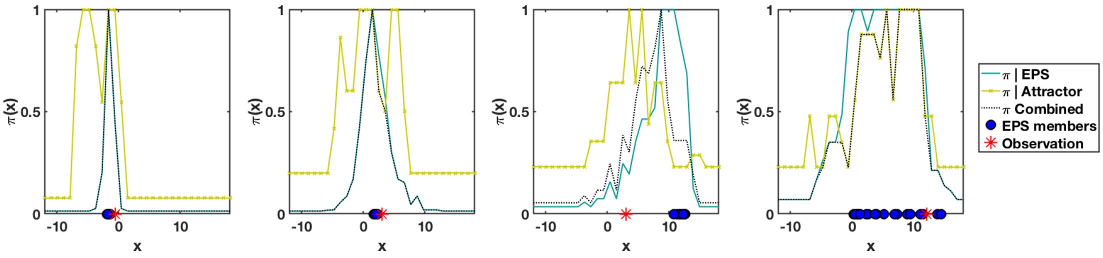


**Figure 4.5** EPS- and attractor-based possibility distributions and their combination at lead times $t = \{1, 3, 5, 7\}$ days (left to right).

the time delay vectors $\boldsymbol{x}_t = \left(x_t, x_{t-\tau}, \ldots, x_{t-(m-1)\tau}\right)$ 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 $(m, \tau)$ 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.

#### 4.3.3.2 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 4.2.2 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.

#### 4.3.3.3 Combining EPS and dynamical information

$\pi_{EPS}$ and $\pi_{DYN}$ 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 their fuzzy intersection [Zadeh, 1978, Hose and Hanss, 2019] (hereafter Zadeh's aggregation):

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

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

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

The resulting distribution is finally normalized to one, to verify axiom 1 from Section 4.2.1. This consists in using the following transformation, for a generic possibility distribution $\pi(x)$:

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

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). This choice can be discussed, for instance in the situation of dependence between the sources of information (see Hose and Hanss [2019] and discussion in Section 4.5.3.1. One may also decide based on additional information (e.g. physics-based, expert opinion, etc), if the two distributions do not overlap at all, to favor one distribution and dismiss the second, making the final distribution less conservative than pure ignorance (but possibly not consistent). 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. abovementioned axioms of definition), these areas are associated to a possibility measure of 1 and other events scaled accordingly. An illustrative example is provided Figure 4.5.

#### 4.3.3.4 Guarantees

We conclude this section with a focus on the formal guarantees that our methodology provides. By construction, the possibility distributions $\pi_{EPS}$ and $\pi_{DYN}$ dominate with a given confidence level $\beta$ (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], $\pi_{COMB}$ still provides guarantees when it comes to the lower bound of $\Pi(A)$. Indeed, from axiom a. of Section 4.2.1, 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 $\pi_{EPS}$ and $\pi_{DYN}$ allows to reduce the basis level $\gamma$ 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 $A$. 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 [2021].

## 4.4 Experimental setting

### 4.4.1 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:

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

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

where $j = 1, \ldots J$ and $k = 1, \ldots K$. 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 $4$th-order method with time step $dt = 0.002$ (model time units) until enough trajectories of duration $1.4$, starting

[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.

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 $X_j$:

$$0.262 - 1.262X_j + 0.004608X_j^2 + 0.007496X_j^3 - 0.0003226X_j^4 \qquad \text{(Equation 8)}$$

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 $X_1$.

### 4.4.2 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 $\tilde{\boldsymbol{x}}_{\boldsymbol{t_0+t}} = \{\tilde{x}^1_{t_0+t}, \dots, \tilde{x}^m_{t_0+t}\}$ is the ensemble, $\theta$ 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 $\pi_{EPS}$, $\pi_{DYN}$ or $\pi_{COMB}$, 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 $\mu$ 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.

### 4.4.3 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, 2021] 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 used 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.

#### 4.4.3.1 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 $\alpha$-cut $C_\pi^\alpha = \{x \in \mathcal{X} \mid \pi(x) \geq \alpha\}$. Another way to compute it in a discretised setting is to order the possibility profile $\pi$ 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 $\pi$ over domain $\mathcal{X}$, obtained respectively for a Dirac-like profile and a uniform profile. Given two possibility profiles $\pi$ and $\pi'$, $U(\pi) \leq U(\pi')$ is equivalent to say that $\pi$ is more specific (i.e. more informative) than $\pi'$.

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

#### 4.4.3.2 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 do 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 4.2.1, $N(A) > 0$ 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 $\pi$ 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:

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

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 $S_{N_\pi}$ 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 $\big(N, \Pi\big)$ to a probability $G$. Beyond the classical $G(O) = \alpha N(O) + (1-\alpha)\Pi(O) = P_\alpha(O)$, where $\alpha$ 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.

#### 4.4.3.3 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.

#### 4.4.3.4 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 $\alpha$ 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 $\alpha_i$-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 $N(A) \leq P(A) \leq \Pi(A)$.

## 4.5 Results and 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

[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 $\approx 2$ 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.

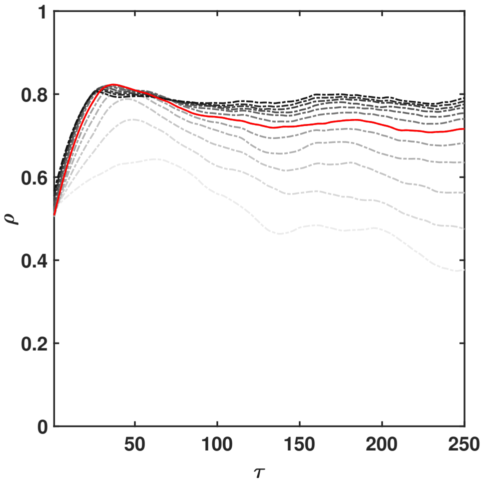


**Figure 4.6** Results of the simplex method applied to the L96 system. The Pearson correlation coefficient between verification at lead time $t = 1$ day and the prediction computed by means of a weighted mean of the $m + 1$ closest analogues in the reconstructed phase space of embedding dimension $m$ and time-delay $\tau$. 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 $\tau$: $m = 9$.

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 $\beta$ 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 $\beta$. Even for small sample sizes, the choice of $\beta$ is not critical and quasi perfect coverage rate is obtained: $\beta \geq 0.8$, ensures that $P\big(P(A) \leq \Pi(A)\big) \rightarrow 1 \quad \forall A$. We consequently use $\beta = 0.9$ which, without impairing guarantees, tends to provide less conservative distributions as shown for the same case study in Le Carrer [2021].

### 4.5.1 Attractor reconstruction

The simplex method introduced in Section 4.3.3 is applied to the lead time $t = 1$ day from the continuous archive $x_{t_1}, \dots, x_{t_{N_{I_A}}}$ 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 4.6). Herafter, when $m$ is not explicitly mentioned for methodologies COMB-$m$ or DYN-$m$, the reader will understand that $m = 9$.

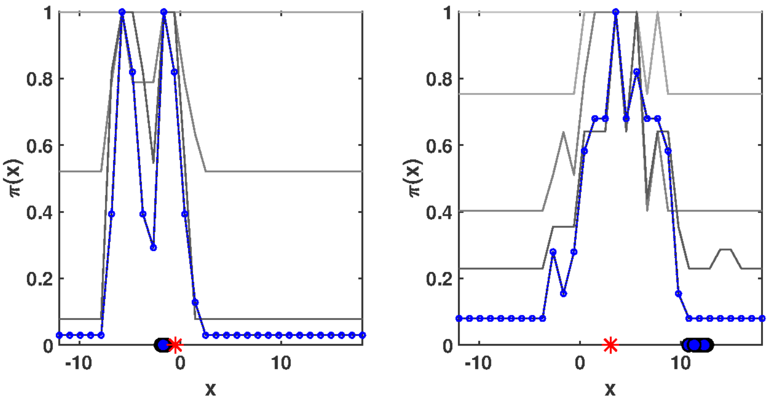


**Figure 4.7** Effect of varying the number of analogs $n_A \in \{10, 50, 100, 500, 1000\}$ (darker lines for larger $n_A$) for lead times $t \in \{1, 5\}$ days (from left to right) on $\pi_{DYN}$. The distribution is highlighted with dots and color for $n_A = 1000$. The smaller $n_A$, the more conservative the distribution. Associated EPS members are marked as blue dots and the verification as a red star.

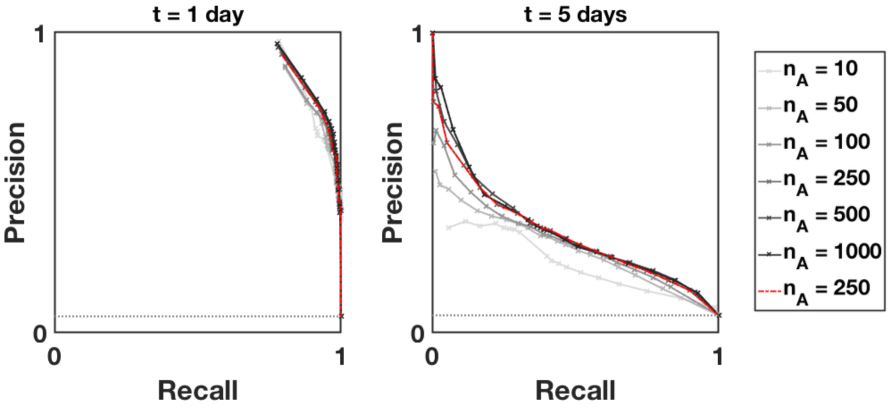


**Figure 4.8** Effect of varying the number of analogs $n_A = \{10, 50, 100, 250, 500, 1000\}$ on the precision-recall curves at lead times 1 and 5 days. The darker the line, the higher $n_A$.

### 4.5.2 Setting the number of analogs $n_A$

As illustrated in Figure 4.7, the parameter $n_A$ plays an important part in the shape of $\pi_{DYN}$ and a careful calibration is consequently recommended. Figure 4.7 shows the effect of increasing the number of analogs $n_A \in \{10, 50, 100, 500, 1000\}$ on $\pi_{DYN}$. 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 tendencies (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 4.8 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 4.7. We continue our experiments with this value for $n_A$.

### 4.5.3 Predictive performances

#### 4.5.3.1 Information content

Figure 4.9 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 4.3.3.3, 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 variables to be fused. Finally, we compare the results for the possibility-based probabilities derived from the methodology pred-CRED and pred-TENT-NEU, and pred-TENT-AV and pred-TENT-PR with varying $\alpha$. Note that the risk-averse and risk-prone versions of the latter 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. Procedures such as climate blending [Bröcker and Smith, 2008] could be used to make these predictive

[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:

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

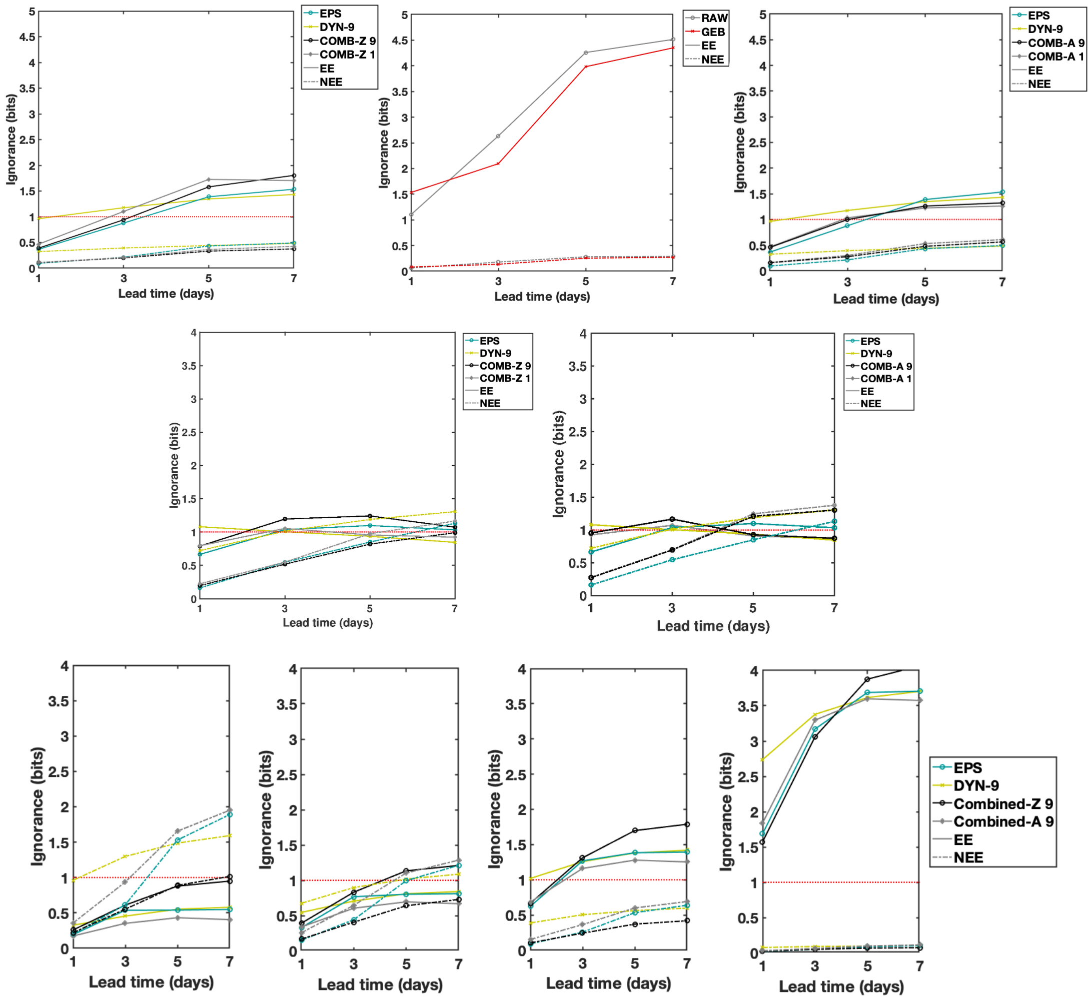


**Figure 4.9** Empirical ignorance score of the methods described in the text. The upper plots use the pred-CRED approach to derive probabilities from possibility distributions. The middle plots use the pred-TENT-NEU and the lower plots use the pred-TENT-$\alpha$ with $\alpha \in \{0.1, 0.25, 0.5, 0.9\}$ 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 cases (top and below), the left-most panels use the COMB-Z aggregation method while the right-most panels use the COMB-A approach for aggregation.

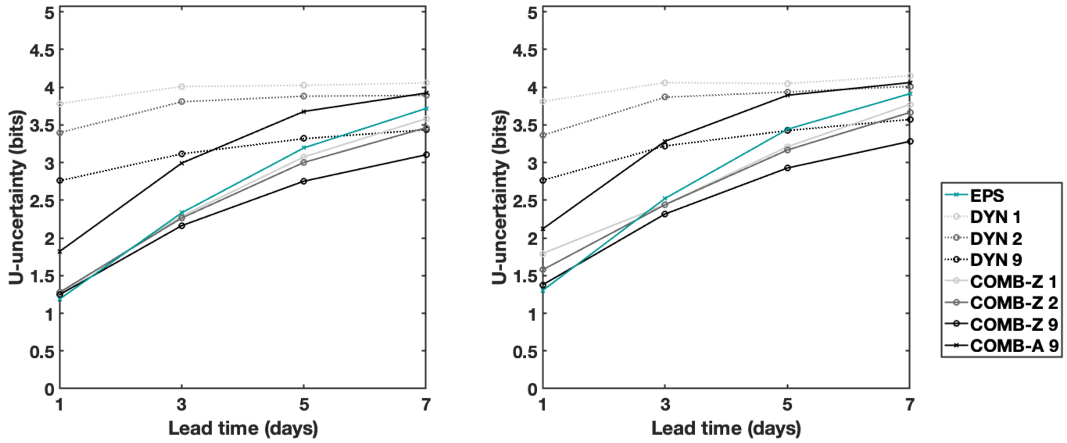


**Figure 4.10** Average U-uncertainty of the possibility distributions described in the text for both NEE (left) and EE (right). The upper bound given the domain of definition of the variable at hand is $\log_2 |\mathcal{X}| = 5.08$, which would be obtained for a uniform possibility distribution.

probabilities more robust to such pitfall, however this is not the point of this paper hence the description of the ignorance score limited to the risk neutral pred-TENT approach.

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 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 4.10 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 4.10, 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 4.9 shows the effect of varying $\alpha$ 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 $\alpha$ 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 $\alpha$ 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.

#### 4.5.3.2 Ability to discriminate

Figure 4.11 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 $\Pi$ doesn't allow small hit rates. Intermediate pooling such as $P_{\alpha=0.5}$ allows to cover the whole range of hit rates. Overall, $\pi_{EPS}$ 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, $\pi_{DYN}$ 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.

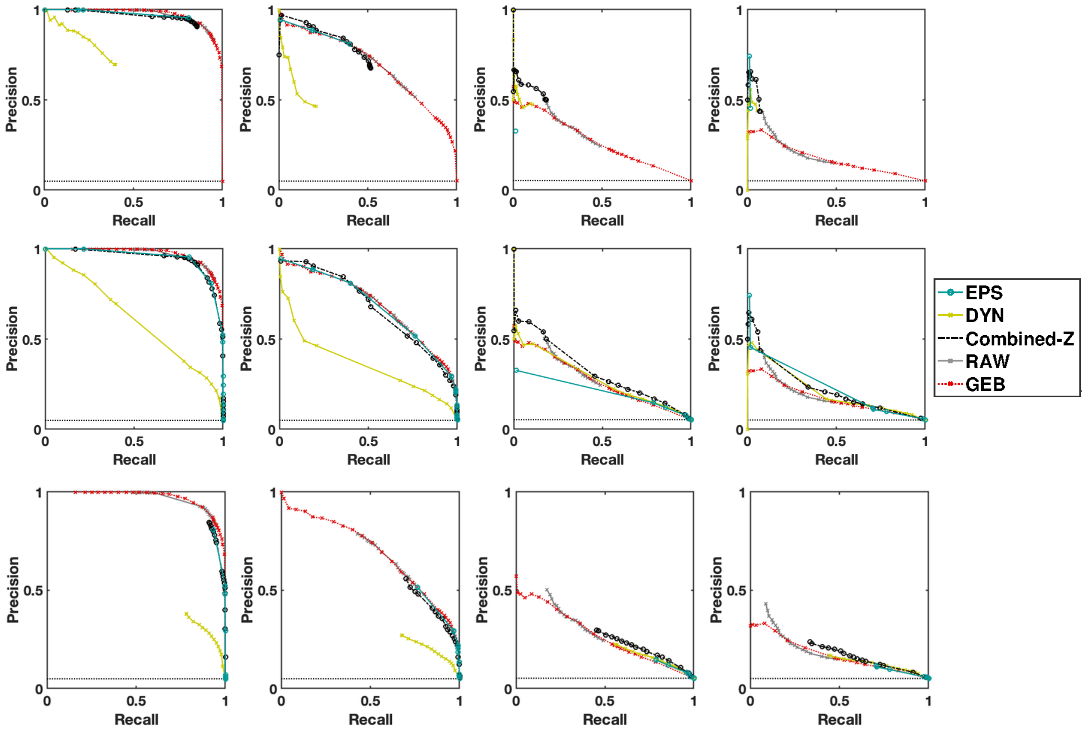


**Figure 4.11** Precision-recall curves showing the predictive skills of possibility distributions EPS, DYN and COMB-Z-9 and probability distributions RAW and GEB for lead times $t = \{1, 3, 5, 7\}$ days (left to right). For the curves associated with the possibilistic approaches, we use $N(A)$, the credibility $0.5\big(N(A) + \Pi(A)\big)$ and $\Pi(A)$ (from top to bottom) as input probabilities.

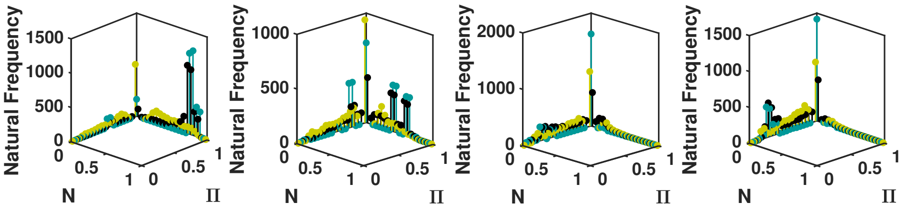


**Figure 4.12** 3-dimensional histograms of the possibilistic predictions associated to verification of $A$ for lead times $t = \{1, 3, 5, 7\}$ days (left to right). Predictions are based on $\pi_{EPS}$ (blue), $\pi_{DYN}$ (yellow) or their Z-combination (black).

We note than the performance of $\pi_{COMB}$ is different than the performance of its best component (either $\pi_{EPS}$ or $\pi_{DYN}$). At small lead times, it remains close to $\pi_{EPS}$ 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 4.5. For short lead times, $\pi_{EPS}$ 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 $\pi_{DYN}$, the latter are often close to the ignorance point as shown by the histogram of the predictions associated with observed events $A$ in Figure 4.12. 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 $t = 5$ 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, $\pi_{EPS}$ 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 $\pi_{DYN}$ through the AND operator allows for a narrower final distribution (the peak at the ignorance point of $\pi_{COMB}$ is smaller in amplitude than the peaks of its components $\pi_{EPS}$ and $\pi_{DYN}$) 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

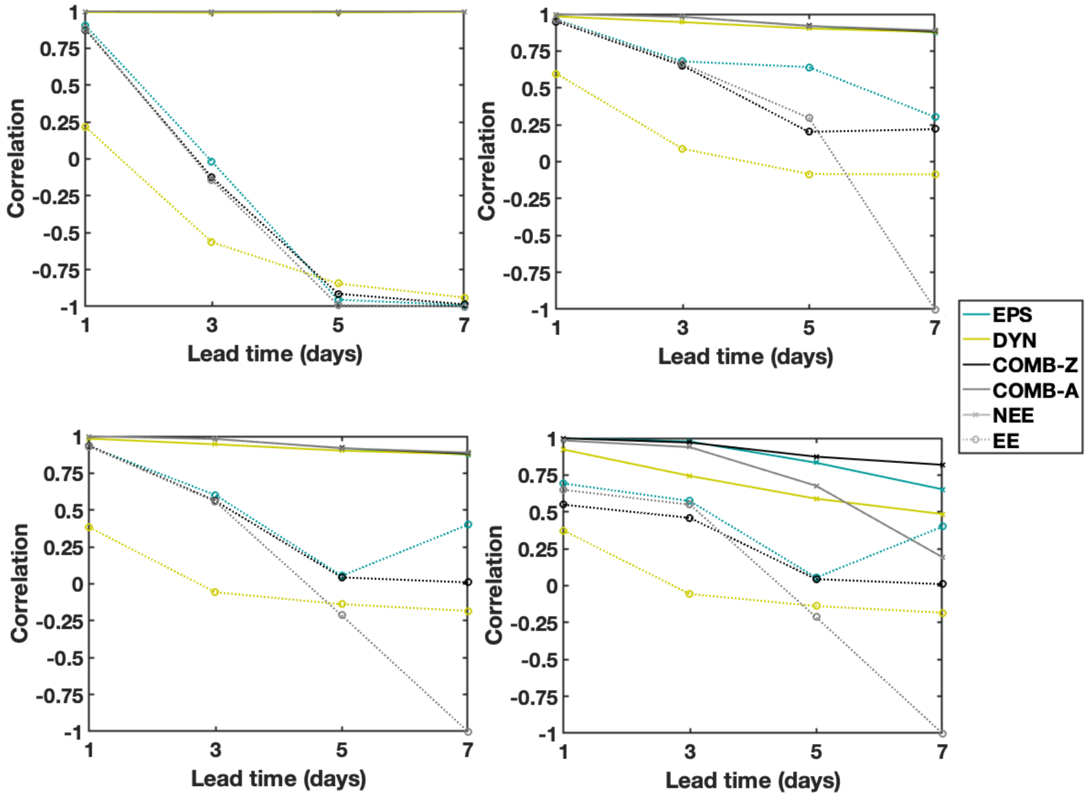


**Figure 4.13** Spearman correlation coefficient between the *a posteriori* propabilistic ignorance score and the level of *a priori* possibilistic ignorance. Results are broken down for EE and NEE, observed (top left) or predicted (others). From left to right and top to bottom, methodologies used are 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.

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.

#### 4.5.3.3 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 $\Pi(A) = N(A) = 0$ ($\bar{A}$ is 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 $W$ (*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 4.13 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, $\pi_{EPS}$, $\pi_{DYN}$ and $\pi_{COMB}$). Besides, to highlight results, we 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 (called NEE).

Figure 4.13 reports the correlation between $S_i$ associated with the probabilities derived from possibilistic methodologies and the associated $W$. It would not make sense to directly compare probabilities from GEB or RAW and the possibilistic $W$ as the latter are issued from different methodologies.

If we break down results between EE and NEE, we observe that possibilistic ($W$) 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 as 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.

In the risk-averse option, results do not change for NEE predicted: the correlation is still very strong and $W$ 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 $W$ 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 $W$ 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 $W$ 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.

#### 4.5.3.4 Reliability

Figure 4.14 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 4.4.3.4). 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.

For all lead times, the envelop of the fuzzy reliability plots covers almost the whole

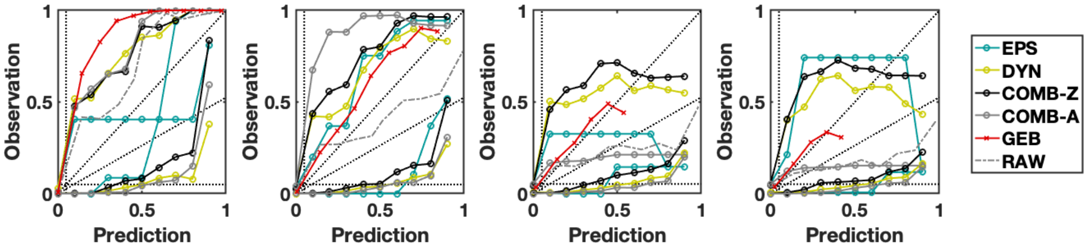


**Figure 4.14** Reliability diagram at lead times $t = \{1, 3, 5, 7\}$ days (left to right). The probabilistic results GEB and RAW are reported in cross-red and dashed grey lines respectively, while the upper and lower bounds of the possibilistic methodologies are in solid-circled lines. Standards elements 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).

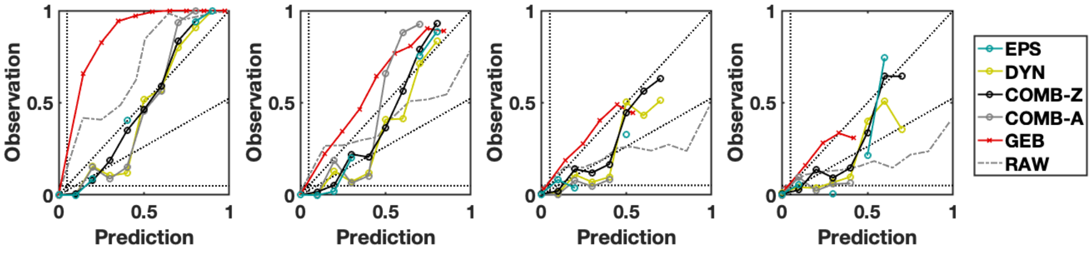


**Figure 4.15** See legend of Figure 4.14. For possibilistic methodologies, we now extract the credibility and use it as a probability to draw the associated reliability diagrams.

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 4.15 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

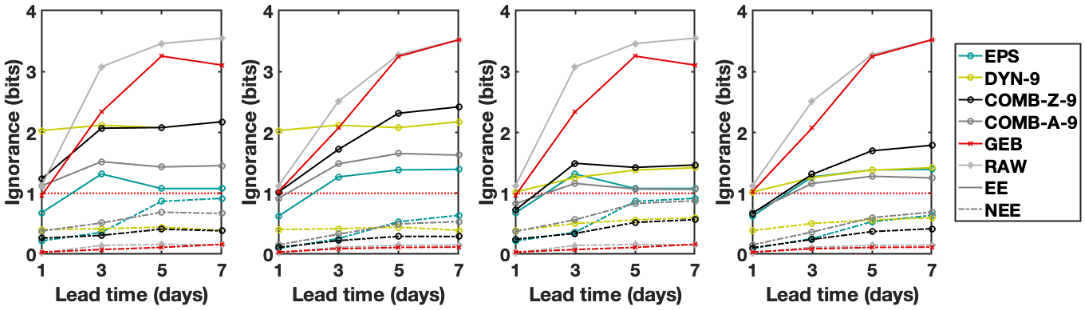


**Figure 4.16** See legend of Figure 4.9. The method used is pred-CRED. The left two diagrams are based on a time series of length about 6 months and the right two ones on a time series of length about 55 years. Within each block of two, we increase from left to right the EPS archive size from 3 years to 30 years.

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 4.14.

#### 4.5.3.5 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 [2021], 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 4.16 presents the empirical ignorance score similarly to Figure 4.9, 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 [2021], we develop this counter-intuitive observation and explain how this is due to the limit behaviour of the possibilistic transformation presented in Section 4.2.1. More points tend to lower the level $\gamma$ 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 $\pi_{DYN}$ the same way it does for $\pi_{EPS}$. 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.

## 4.6 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.

## 4.7 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 $x^i$: $B = \{b_i, i = 1, \dots, n\}$ 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 4.8, 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 $n_i > 5, i = 1, \dots, n$ 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 $\{\sigma_l, l = 1, \dots, L\}$, where $\sigma_l$ is the permutation of the indices $\{1, \dots, n\}$ associated to $\mathcal{P}$ such that $p^+_{\sigma_l(1)} < p^-_{\sigma_l(2)}, p^+_{\sigma_l(2)} < p^-_{\sigma_l(3)}, \dots, p^+_{\sigma_l(n-1)} < p^-_{\sigma_l(n)}$ or equivalently $\sigma_l(i) < \sigma_l(j) \Leftrightarrow \left(b_{\sigma(i)}, b_{\sigma(j)}\right) \in \mathcal{P}$. $\sigma$ is a bijection and the reverse transformation $\sigma^{-1}$ 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 $\sigma_l$ and each class $b_i$, solve the following linear program:

$$\pi_i^{\sigma_l} = \max_{p1,\dots,p_n} \sum_{j | \sigma_l^{-1}(j) \leq \sigma_l^{-1}(i)} p_j \qquad \text{(Equation 15)}$$

under the constraints

$$\begin{cases} \sum_{k=1}^{K} p_k = 1 \\ p_k^- \leq p_k \leq p_k^+ \quad \forall k \in \{1, \dots, n\} \\ p_{\sigma_l(1)} \leq p_{\sigma_l(2)} \leq \dots \leq p_{\sigma_l(n)} \quad . \end{cases} \tag{Equation 16}$$

5. Finally, take the distribution dominating all the distributions $\pi^{\sigma_l}$:

$$\pi_i = \max_{l=1,\dots,L} \pi_i^{\sigma_l} \; \forall i \in \{1, \dots, n\}. \tag{Equation 17}$$

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.

## 4.8 Goodman [1965]'s formulation

Following the problem and notation introduced in Appendix 4.7, if we note:

$$A = \chi^2(1-\beta/n, 1) + N_s \;, \tag{Equation 18}$$

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,

$$B_i = \chi^2(1-\beta/n, 1) + 2n_i \;, \tag{Equation 19}$$

$$C_i = B_i^2 - 4AC_i \;, \tag{Equation 20}$$

$$\Delta_i = \frac{n_i^2}{N_s} \;, \tag{Equation 21}$$

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 $b_i$, $i = 1, \dots, n$ are given by:

$$[p_i^-, p_i^+] = \left[\frac{B_i - \sqrt{\Delta_i}}{2A}, \frac{B_i + \sqrt{\Delta_i}}{2A}\right] . \tag{Equation 22}$$

# Bibliography

Sam Allen, Christopher AT Ferro, and Frank Kwasniok. Regime-dependent statistical post-processing of ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*, 145(725):3535–3552, 2019.

Jochen Bröcker and Leonard A. Smith. Increasing the Reliability of Reliability Diagrams. *Weather and Forecasting*, 22(3):651–661, 2007. doi: 10.1175/WAF993.1.

Jochen Bröcker and Leonard A. Smith. From ensemble forecasts to predictive distribution functions. *Tellus A: Dynamic Meteorology and Oceanography*, 60(4):663–678, 2008. doi: 10.1111/j.1600-0870.2007.00333.x.

Roberto Buizza. Ensemble Forecasting and the Need for Calibration. In *Statistical Postprocessing of Ensemble Forecasts*, pages 15–48. Elsevier, 2018. ISBN 978-0-12-812372-0. doi: 10.1016/B978-0-12-812372-0.00002-9.

Roberto Buizza, M Milleer, and Tim N Palmer. Stochastic representation of model uncertainties in the ECMWF ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society*, 125(560):2887–2908, 1999. doi: 10.1002/qj.49712556006.

Aurélien Ben Daoud, Eric Sauquet, Guillaume Bontron, Charles Obled, and Michel Lang. Daily quantitative precipitation forecasts based on the analogue method: Improvements and application to a French large river basin. *Atmospheric Research*, 169: 147–159, 2016. ISSN 0169-8095. doi: 10.1016/j.atmosres.2015.09.015.

Gert De Cooman and Dirk Aeyels. Supremum preserving upper probabilities. *Information Sciences*, 118(1):173–212, 1999. ISSN 0020-0255. doi: https://doi.org/10.1016/S0020-0255(99)00007-9.

M. Delgado and S. Moral. On the concept of possibility-probability consistency. *Fuzzy Sets and Systems*, 21(3):311–318, 1987. ISSN 0165-0114. doi: https://doi.org/10.1016/0165-0114(87)90132-1.

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

Didier Dubois and Henri Prade. On several representations of uncertain body of evidence. *Fuzzy Information and Decision Processes*, pages 167–181, 1982.

Didier Dubois and Henri Prade. *Possibility theory: an approach to computerized processing of uncertainty*. Springer Science and Business Media, 2012.

Didier Dubois and Henry Prade. Possibility theory and its applications: Where do we stand? In *Springer handbook of computational intelligence*, pages 31–60. Springer, 2015.

Didier Dubois, Henri Prade, and Sandra Sandri. On Possibility/Probability Transformations. In R. Lowen and M. Roubens, editors, *Fuzzy Logic: State of the Art*, pages 103–112. Springer Netherlands, Dordrecht, 1993. ISBN 978-94-011-2014-2. doi: 10.1007/978-94-011-2014-2_10.

Didier Dubois, Laurent Foulloy, Gilles Mauris, and Henri Prade. Probability-Possibility Transformations, Triangular Fuzzy Sets, and Probabilistic Inequalities. *Reliable computing*, 10(4):273–297, 2004. doi: 10.1023/B:REOM.0000032115.22510.b5.

Didier Dubois, Henri Prade, and Philippe Smets. A definition of subjective possibility. *International journal of approximate reasoning*, 48(2):352–364, 2008.

P. Friederichs and A. Hense. Statistical Downscaling of Extreme Precipitation Events Using Censored Quantile Regression. *Monthly Weather Review*, 135(6):2365–2378, 2007. doi: 10.1175/MWR3403.1.

Petra Friederichs, Sabrina Wahl, and Sebastian Buschow. Postprocessing for Extreme Events. In Stéphane Vannitsem, Daniel S. Wilks, and Jakob W. Messner, editors, *Statistical Postprocessing of Ensemble Forecasts*, pages 127–154. Elsevier, 2018. ISBN 978-0-12-812372-0. doi: https://doi.org/10.1016/B978-0-12-812372-0.00005-4.

Tilmann Gneiting and Matthias Katzfuss. Probabilistic forecasting. *Annual Review of Statistics and Its Application*, 1:125–151, 2014.

Tilmann Gneiting, Adrian E. Raftery, Anton H. Westveld, and Tom Goldman. Calibrated Probabilistic Forecasting Using Ensemble Model Output Statistics and Minimum CRPS Estimation. *Monthly Weather Review*, 133(5):1098–1118, 2005. doi: 10.1175/MWR2904.1.

Irving John Good. How to Estimate Probabilities. *IMA Journal of Applied Mathematics*, 2(4):364–383, 12 1966. ISSN 0272-4960. doi: 10.1093/imamat/2.4.364.

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

Carlo Graziani, Robert Rosner, Jennifer M Adams, and Reason L Machete. Probabilistic Recalibration of Forecasts. *arXiv preprint arXiv:1904.02855*, 2019.

Renate Hagedorn, Thomas M Hamill, and Jeffrey S Whitaker. Probabilistic forecast calibration using ecmwf and gfs ensemble reforecasts. part i: Two-meter temperatures. *Monthly Weather Review*, 136(7):2608–2619, 2008.

Thomas M Hamill and Stephen J Colucci. Verification of Eta–RSM Short-Range Ensemble Forecasts. *Monthly Weather Review*, 125(6):1312–1327, 1997.

Thomas M. Hamill and Michael Scheuerer. Probabilistic precipitation forecast postprocessing using quantile mapping and rank-weighted best-member dressing. *Monthly Weather Review*, 146(12):4079–4098, 2018. doi: 10.1175/MWR-D-18-0147.1.

Thomas M. Hamill and Jeffrey S. Whitaker. Probabilistic Quantitative Precipitation Forecasts Based on Reforecast Analogs: Theory and Application. *Monthly Weather Review*, 134(11):3209–3229, 2006. doi: 10.1175/MWR3237.1.

Thomas M Hamill, Jeffrey S Whitaker, and Xue Wei. Ensemble reforecasting: Improving medium-range forecast skill using retrospective forecasts. *Monthly Weather Review*, 132(6):1434–1447, 2004.

D. L. Hartmann, R. Buizza, and T. N. Palmer. Singular Vectors: The Effect of Spatial Scale on Linear Growth of Disturbances. *Journal of the Atmospheric Sciences*, 52(22): 3885–3894, 1995. doi: 10.1175/1520-0469(1995)052<3885:SVTEOS>2.0.CO;2.

Dominik Hose and Michael Hanss. Possibilistic calculus as a conservative counterpart to probabilistic calculus. *Mechanical Systems and Signal Processing*, 133:106290, 2019.

George J Klir. Uncertainty and information. *Foundations of Generalized Information Theory*, 2006.

Noémie Le Carrer. Possibly extreme, probably not: Is possibility theory the route for risk-averse decision-making? *Atmospheric Science Letters*, page e01030, 2021.

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

T. P. Legg and K. R. Mylne. Early Warnings of Severe Weather from Ensemble Forecast Information. *Weather and Forecasting*, 19(5):891–906, 2004. doi: 10.1175/1520-0434(2004)019<0891:EWOSWF>2.0.CO;2.

C. E. Leith. Theoretical Skill of Monte Carlo Forecasts. *Monthly Weather Review*, 102 (6):409–418, 1974. doi: 10.1175/1520-0493(1974)102<0409:TSOMCF>2.0.CO;2.

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

Edward N Lorenz. Empirical orthogonal functions and statistical weather prediction. 1956.

Edward N. Lorenz. Atmospheric Predictability as Revealed by Naturally Occurring Analogues. *Journal of the Atmospheric Sciences*, 26(4):636–646, 1969. doi: 10.1175/1520-0469(1969)26<636:APARBN>2.0.CO;2.

Edward N Lorenz. Predictability: A problem partly solved. In *Proc. Seminar on predictability*, volume 1, 1996.

Jiayi Ma, Ming Yang, Xueshan Han, and Zhi Li. Ultra-Short-Term Wind Generation Forecast Based on Multivariate Empirical Dynamic Modeling. *IEEE Transactions on Industry Applications*, 54(2):1029–1038, 2017.

Marie-Hélène Masson and Thierry Denœux. Inferring a possibility distribution from empirical data. *Fuzzy Sets and Systems*, 157(3):319–340, 2006. ISSN 0165-0114. doi: https://doi.org/10.1016/j.fss.2005.07.007.

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

K Mylne, C Woolcock, J Denholm-Price, and R Darvell. 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*, pages 113–118, 2002.

David Orrell. Ensemble Forecasting in a System with Model Error. *Journal of the Atmospheric Sciences*, 62(5):1652–1659, 2005. doi: 10.1175/JAS3406.1.

Paul Platzer, Pascal Yiou, Philippe Naveau, Pierre Tandeo, Yicun Zhen, Pierre Ailliot, and Jean-François Filipot. Using local dynamics to explain analog forecasting of chaotic systems. *Journal of the Atmospheric Sciences*, 2021.

Adrian E. Raftery, Tilmann Gneiting, Fadoua Balabdaoui, and Michael Polakowski. Using Bayesian Model Averaging to Calibrate Forecast Ensembles. *Monthly Weather Review*, 133(5):1155–1174, 2005. doi: 10.1175/MWR2906.1.

Nandini Ramesh and Mark A. Cane. The Predictability of Tropical Pacific Decadal Variability: Insights from Attractor Reconstruction. *Journal of the Atmospheric Sciences*, 76(3):801–819, 2019. doi: 10.1175/JAS-D-18-0114.1.

Mark S. Roulston and Leonard A. Smith. Evaluating Probabilistic Forecasts Using Information Theory. *Monthly Weather Review*, 130(6):1653–1660, 2002. doi: 10.1175/1520-0493(2002)130<1653:EPFUIT>2.0.CO;2.

Mark S. Roulston and Leonard A. Smith. Combining dynamical and statistical ensembles. *Tellus A: Dynamic Meteorology and Oceanography*, 55(1):16–30, 2003. doi: 10.3402/tellusa.v55i1.12082.

Takaya Saito and Marc Rehmsmeier. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PloS one*, 10(3), 2015.

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

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

Leonard A. Smith. Integrating Information, Misinformation and Desire: Improved Weather-Risk Management for the Energy Sector. In Philip J. Aston, Anthony J. Mulholland, and Katherine M.M. Tant, editors, *UK Success Stories in Industrial Mathematics*, pages 289–296. Springer International Publishing, Cham, 2016. ISBN 978-3-319-25454-8. doi: 10.1007/978-3-319-25454-8_37.

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

George Sugihara, Robert May, Hao Ye, Chih-hao Hsieh, Ethan Deyle, Michael Fogarty, and Stephan Munch. Detecting Causality in Complex Ecosystems. *Science*, 338(6106): 496–500, 2012. ISSN 0036-8075. doi: 10.1126/science.1227079.

Floris Takens. Detecting strange attractors in turbulence. In David Rand and Lai-Sang Young, editors, *Dynamical Systems and Turbulence, Warwick 1980*, pages 366–381. Springer Berlin Heidelberg, Berlin, Heidelberg, 1981. ISBN 978-3-540-38945-3.

Zoltan Toth and Eugenia Kalnay. Ensemble Forecasting at NCEP and the Breeding Method. *Monthly Weather Review*, 125(12):3297–3319, 1997. doi: 10.1175/1520-0493(1997)125<3297:EFANAT>2.0.CO;2.

Anna Trevisan. Statistical Properties of Predictability from Atmospheric Analogs and the Existence of Multiple Flow Regimes. *Journal of the Atmospheric Sciences*, 52(20): 3577–3592, 1995. doi: 10.1175/1520-0469(1995)052<3577:SPOPFA>2.0.CO;2.

H.M. Van den Dool. Searching for analogues, how long must we wait? *Tellus A*, 46(3): 314–324, 1994. doi: 10.1034/j.1600-0870.1994.t01-2-00006.x.

Peter Walley. Inferences from Multinomial Data: Learning About a Bag of Marbles. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):3–34, 1996. doi: 10.1111/j.2517-6161.1996.tb02065.x.

Daniel S. Wilks and Thomas M. Hamill. Potential Economic Value of Ensemble-Based Surface Weather Forecasts. *Monthly Weather Review*, 123(12):3565–3575, 1995. doi: 10.1175/1520-0493(1995)123<3565:PEVOEB>2.0.CO;2.

R. M. Williams, C. A. T. Ferro, and F. Kwasniok. A comparison of ensemble post-processing methods for extreme events. *Quarterly Journal of the Royal Meteorological Society*, 140(680):1112–1120, 2014. doi: 10.1002/qj.2198.

L.A Zadeh. Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets and Systems*, 1 (1):3–28, 1978. ISSN 0165-0114. doi: https://doi.org/10.1016/0165-0114(78)90029-5.

# Chapter 5

# A possibilistic interpretation of ensemble predictions: Application to shipping optimisation in tidal areas

In this last chapter, we apply the predictive framework developed in Chapters 3 and 4 to a real-world problem. Namely, we go back to the initial problem of tidal ship routing developed in Chapter 2 and study the respective values of the following different sources of sea level information in the optimisation process:

(a) Deterministic tide predictions associated with a safety margin;

(b) Probabilistic modelling of the residuals as performed in Chapter 2;

(c) Ensemble predictions of residuals interpreted according to the possibilistic framework developed in the two previous chapters or according to a classical probabilistic ensemble dressing.

In the possibilistic case, we cannot use the same optimisation algorithm as developed for classical probabilistic residual predictions. Consequently we adapt to our problem a methodology developed by Hose et al. [2018] for global optimisation with possibilistic variables.

Due to the limitation in available data (we had at hand 6 months of 6-days ahead EPS predictions for two ports: Liverpool and Southampton), we restrict our study to a simple $N = 2$ -port case study similar to the one presented in Chapter 2, Section 2.2.1. Beyond its limitations, such an analysis already allows us to draw some trends of results and to draw the possible advantages and drawbacks of our data-driven possibilistic approach in a real-world application, in particular in the case of limited datasets. Beyond that, it is another application of possibility-based global optimisation, which is interesting for highlighting the potential benefits and limitations of such an approach w.r.t. more classical probability-based methodologies.

The contributions of the authors is the following: NLC found the research idea, designed the methodology, the experiments and implemented them, as well as analysed the results. She wrote the paper.

# A possibilistic interpretation of ensemble predictions: Application to shipping optimisation in tidal areas

N. Le Carrer, S. Ferson

Submitted in March 2021 to the *2021 European Safety and Reliability Conference.*

### Abstract

Until now, works in the field of tide routing (i.e., optimization of cargo loading and ship scheduling decisions in tidal ports and shallow seas) have omitted the uncertainty of sea level predictions. However, the widely used harmonic tide forecasts are not perfectly reliable. Consequences for the maritime industry are significant: current solutions to tide routing may be made robust through the introduction of arbitrary slack, but they are not optimal [Le Carrer et al., 2020]. Given the financial implications at stake for every additional centimeter of draft and the catastrophic effects of a grounding, an investigation of tide routing from the perspective of risk analysis is necessary.

Ensemble forecasting has gained popularity in the field of numerical weather prediction as a way of quantifying the uncertainty on forecasts. Tide-surge ensemble forecasts are routinely produced as well, combining hydrodynamic models with weather ensembles. This type of forecasts is commonly interpreted in a probabilistic way. However, the latter is regularly criticized for not being reliable, especially for predicting extreme events because of the chaotic nature of the dynamics of the atmospheric-ocean system, model error, and the fact that ensemble of forecasts are not, in reality, produced in a probabilistic manner. In this work, we develop a possibilistic framework to interpret and use operationally such ensembles of predictions.

Considering the journey of a bulk carrier between a set of ports, a shipping decision model is designed to compute optimal cargo loading and scheduling decisions, given the time series of the fuzzy sea levels in these ports. The under-keel clearance becomes a fuzzy constraint and the resulting shipping optimization problem is solved by means of the possibilistic approach developed by Hose et al. [2018]. Results obtained on a realistic case study with 6-day-ahead tide surge ensemble predictions are discussed and compared with those given by a probabilistic approach, or by standard practices on ships. They illustrate the potential and limitations of a possibilistic interpretation of the weather ensemble forecasts over its probabilistic counterpart in a realistic setting.

**Keywords** — Robust Optimization, OR in maritime industry, Uncertainty modeling, Decision making, Uncertainty propagation, Possibility theory

## 5.1 Introduction

### 5.1.1 Robust maritime shipping optimisation in tidal areas

In spite of its significance for global trade, maritime shipping remains an activity constrained by a range of uncertain factors [Goerlandt and Montewka, 2015, Song and Furman, 2013]. Beyond the well-known weather at sea (e.g. [Azaron and Kianfar, 2003]), berth occupation (e.g. [Agra et al., 2015]), dockers availability, bunker fuel prices or market demand (e.g. [Chuang et al., 2010]), sea levels in shallow waters can impact significantly the outputs of a maritime shipping operation. Knowing that an extra centimetre of draft corresponds approximately to 50 tons of cargo for an average bulk carrier [Uslu et al., 2017], being able to predict accurately water levels in ports translates into economic benefits for both shipping operators (economies of scale) and port authorities (vessel throughput). Deterministic harmonic tide predictions are traditionally used to estimate the future sea levels in shallow waters. From these ones, a shipper can estimate how much freight to load in order to ensure a positive under-keel clearance (UKC; the distance between the deepest underwater point of the ship and the seabed), which includes a safety margin dictated from authorities. Yet sea levels are impacted by environmental factors (wind, pressure, currents) that locally increase or decrease the actual sea levels w.r.t. the harmonic predictions. The difference, hereafter residual, can be significant. Thus, overall British tide stations, residuals are typically 10 cm and rise to 29 cm for high tidal range stations [Flowerdew et al., 2010]. Similarly, sea level residuals can amount to 30% of the total measured sea level in Hillarys Boat Harbour, Western Australia Makarynskyy et al. [2004]. Whether to load more, depart earlier, or catch a tide window, recent works have shown the economical value of modeling sea level residuals beyond a traditional 'rule-of-the-thumb' safety margin on tide predictions.

This trend started with the work of Kelareva [2011], Kelareva et al. [2012], who developed the concept of dynamic UKC to optimise ship scheduling and cargo loading decisions of multiple vessels at a single port. To estimate the dynamic UKC, the authors deduct from the port depth and predicted tide, not only the vessel's draft, but also a number of allowances accounting for the dynamical responses of the hull to its environment (squat, heeling, wave, water density variation), the tidal prediction error and the variability of bathymetry [Galor, 2008]. Kelareva [2011] use short-term predictions of the dynamic under-keel clearance provided by the DUKC® software (OMC International, 1993, described in Kelareva et al. [2012], O'Brien et al. [2002]). Specifically, from real-time environmental measurements (water depths, wind, waves, current) and ship information (trim, speed, acceleration), the physical responses to the ship moving in a dynamic environment are computed and the dynamic under-keel clearance is estimated. The optimal cargo loading and short term ship scheduling decisions, given this estimation, are then computed. Such a solution is based on real-time measurement of the sea state and provides under-keel clearance information for the upcoming tide-window only [Kelareva et al., 2012]. The economic gains of such a dynamic modelling the UKC are documented in a range of case studies, for optimising cargo load and port throughput [O'Brien et al., 2002] or berth-to-berth voyage scheduling optimisation [Hibbert et al., 2019]. The DUKC®'s short term UKC predictions are now informed by sea level predictions from two distinct models: a global oceanic "weather" model (coastal currents, mesoscale eddies, etc) and a refined sea level model at the port scale [Uslu et al., 2017]. Both are assimilated by means of a Bayesian recursive approach, where residual are assumed Gaussian, allowing improved 7−day ahead predictions for operational use. In parallel, [Le Carrer et al., 2020] showed on a numerical case study how a simple best-fit stochastic modelling of the tide residuals in each port of call allowed to improve the robustness of loading and scheduling decisions and the corresponding net benefits for planning horizons of a week.

### 5.1.2 Tidal residual ensemble predictions

In the early 2000s, the field of weather forecasting saw an important shift in paradigm. While for the past fifty years, improving forecast skills translated into improving the numerical, deterministic model resolution, ensembles of lower resolution predictions became the new norm to forecast the future weather and assess in terms of probabilities the uncertainty on such predictions [Palmer, 2019]. This came from the realisation that weather forecasts are limited, in addition to the numerical representation of physical processes and resolution of the simulations, by the sensitivity of the solutions to the initial conditions and sub-grid parameterisation [Buizza, 2018]. Given sufficient computational resources, it became interesting to sample a limited number (typically 10-50) of initial conditions and then run the numerical weather model, possibly stochastically parameterised, for each of them. The corresponding set of predictions for a given place and lead time, can, after post-processing, be interpreted in probabilistic ways, as the probability distribution of the future state of the atmosphere [Gneiting et al., 2005, Gneiting and Katzfuss, 2014].

To mitigate the possibly high impact damages from hydrodynamic processes such as storm surges [Gerritsen, 2005, De Zolt et al., 2006], operational forecasting and warning systems have been set up in the regions affected by such events (e.g. the coastal flood warnings from the Environment Agency in Wales and England). They aim both at driving the improvement of storm surge forecast skills as well as their interpretation, in order to emit early warnings and take the adequate level of protective measures when necessary. The uncertainty associated with the prediction of storm surge is assumed to be dominated by the driving atmospheric forecast of conditions at the sea surface [Flowerdew et al., 2010]. Operational storm surge forecasting systems consequently now consider the $M$ members of regional weather EPS as input for their storm surge prediction model. The hydrodynamic tide-surge model is run for each member of the weather EPS, namely each one of the $M$ fields of sea-level pressure and 10-metre wind speed, which provides an ensemble of $M$ storm surge predictions at a given coastal place and time. The United Kingdom dispose of its medium-range storm surge ensemble [Flowerdew et al., 2013] with weather ensemble inputs from the Met Office Global and Regional Ensemble Prediction System (MOGREPS). Similarly, the Dutch Meteorological Institute runs an operational storm surge ensemble for the Netherlands with weather ensemble inputs provided by the European Centre for Medium-Range Weather Forecasts (ECMWF) [de Vries, 2009]. Mel and Lionello [2014a,b] present and analyse the performances of a storm surge EPS over the northern Adriatic Sea, with input from the ECMWF as well. New York City disposes as well of two storm surge ensembles, described and gathered in a multi-model ensemble in Di Liberto et al. [2011].

Ensemble predictions are particularly attractive as they are designed to issue probabilistic predictions about a given event. This is typically of interest in the case of flood management. From the probability that the water overtops coastal defences, the authorities can proceed with e.g. a cost-loss analysis and assess whether protective measure must be taken or not [Richardson, 2000, Flowerdew et al., 2013]. Flowerdew et al. [2010, 2013] and Mel and Lionello [2014b, 2016] uses the fraction of members that are above the threshold of interest for at least one time step over a given time window (e.g. 12 hours) as an estimate of the probability that the surge crosses this threshold around a given lead time. They both show that such a probability forecast performs better than a deterministic forecast dressed (e.g. by means of the root mean square error) as a probability distribution and has clear predictive skills.

In weather ensemble forecasting, it is acknowledged that by design ensemble forecasts do not provide actionable ’probabilities’, especially for extreme events [Mylne et al., 2002, Gneiting and Katzfuss, 2014], and post-processing is needed to make them reliable and thus operational. Recalibration of ensemble forecasts became a field of research on its own [Buizza, 2018], whose aim can be summarized as "find the transformation that, applied to the raw ensemble, leads to the probability distribution that will maximise a performance metric on the long term, in particular improve their reliability". In previous works [Le Carrer and Green, 2020, Le Carrer, 2021, Le Carrer and Ferson, 2020], the authors questioned the probabilistic interpretation of an EPS and instead considered that the EPS should be seen in a possibilistic way. They used possibility theory [Zadeh, 1978, Dubois and Prade, 2012] to build a framework allowing to interpret more directly EPS and to provide guaranteed bounds on probabilities.

### 5.1.3 Summary of contributions and outline

In this work, we wonder what is the value of storm surge ensemble predictions for maritime shipping optimisation. To that end, we use the ship scheduling and loading problem developed in previous work by the authors [Le Carrer et al., 2020]. The objective was to find the optimal loading/unloading and scheduling decision in each port of call, which leads to the maximal and robust shipping net benefit, given parameters of the journey and harmonic tide predictions as well as best-fit distribution of residuals in each port of call. In this work, we design two procedures to use the ensemble predictions as source of information on the future residuals. One is based on their probabilistic interpretation, and the other comes from their possibilistic interpretation. We compare their respective performances when it comes to the shipping optimisation problem. These are as well assessed against the performance of our previous methodology, based on tide predictions and best-fit modelling of the residuals [Le Carrer et al., 2020]. This allows us to discuss the value of EPS information and of its varied interpretations. Besides, as noted in Le Carrer [2021], Le Carrer and Ferson [2020] and as generally stressed for analog-based methods, the performance of our methodology depends on the size of the archives (EPS and past time series of the variable of interest) at hand. Since we have very small datasets (1 year) for each port, this study also allows us to test the performance of our approach in a real-world problem with limited archives available.

In this work, we discuss the benefits of (i) using a framework based on possibility theory for extracting the information contained in a storm-surge EPS; (ii) combining it with the insight about the local dynamics of the system gained from the analog method; and (iii) using this possibilistic information in a simple maritime shipping optimisation problem. Our investigation is particularly driven by the following questions:

1. How valuable is the information extracted from the storm-surge EPS, either via a probabilistic approach or via a possibilistic approach, for an application such as maritime shipping optimisation?

2. In particular, is this information more valuable for this specific application than a classical Monte-Carlo-based optimisation using harmonic tide predictions and historical best-fit modelling of sea level residuals in each port?

Section 5.2 presents the case study with the data at hand. Section 5.3 summarises the possibilistic framework for EPS interpretation and compares the predictive performances of this framework on the storm-surge data at hand for our case study w.r.t. the predictive performances of a classical probabilistic ensemble dressing. Section 5.4 develops the fuzzy-constrained interpretation of the maritime shipping optimisation problem and shows how to solve it by means of the outputs of the possibilistic interpretation of storm-surge EPSs. Section 5.5 finally presents the experiments and their results, while Section 5.6 recapitulates the main conclusions of this study.

## 5.2 Case study

We consider a simplified version of the maritime inventory routing problem, where a material is produced at a given rate in a production site (the loading port) and consumed at other sites (unloading ports), at specified rates. Given storage capacities in the production and consumption locations, the general inventory routing problem consists in finding the optimal design of routes and fleet schedule that minimises the shipping costs (sailing and port costs) without interrupting any of the production or the consumption in the aforementioned sites. In our case study, we assume that an industrial operator has sea level forecasts at $N$ ports, at a given time $t_0$ and over a prediction horizon $T$. On this basis, the operator has to decide the total amount $m_1$ of a given commodity to load on a general cargo ship at departure port $p_1$, and the fraction of this cargo $m_{j-1} - m_j$ that will be delivered in each of the remaining ports $p_j, j = \{2, \dots, N\}$, as well as the estimated departure times $t_j$ in each port. The deliveries all have to satisfy the constraints of the inventory routing problem, namely to match a given demand $a_j$ in each port. We assume that all ports have unlimited storage capacities. The optimisation is made on an industrial shipping basis. In other words, the shipper owns the material to be shipped and wants to maximise the net benefit $B$ of the shipment (the value of the cargo loaded minus the shipping costs). We assume a constant ship speed $v$, provided by the ship specifications, along the $N-1$ legs of length $l_j, j = 1, \dots, N-1$ between the departure and arrival ports.

### 5.2.1 Objective function

The goal of the optimisation problem is to find the decision vector $\boldsymbol{d} = \big(m_1, t_1, \dots, m_{N-1}, t_{N-1}\big)$ that optimises the net benefit $B(\boldsymbol{d})$, given the vectors $\hat{\boldsymbol{X}}_j$ of sea level predictions $\hat{X}_j(t)$ available at decision time $t_0$ spanning the horizon $t \in [t_0, t_0 + T]$ in the entrance channels of each port $p_j$, given the constraints $a_j$ on the demand and given constraints from the ship design (carrying capacity), safety at sea (minimum acceptable water under keel), port management (opening times and price bands for port labour), that depend on a range of *a priori* fixed parameters gathered in vector $\boldsymbol{\Theta}$. The shipping return is given by:

$$B(\boldsymbol{d}; \boldsymbol{\Theta}, \{\hat{\boldsymbol{X}}_j\}_{j=1,\dots,N}) = \begin{cases} V - (O + P + U) & \text{if delivered on time,} \\ Z & \text{otherwise.} \end{cases}$$

(Equation 1)

$V = C_c.m_1$ is the merchant value of the cargo, where $C_c$ is the unit value of the freight. From there, we subtract the operational costs of the journey, starting from $t_0$ with an empty ship and finishing at $t_a + \frac{m_N}{u_N}$ after unloading the material in port $p_N$ where the ship arrived at time $t_a$ and unloaded the residual freight at speed $u_N$. These charges encompass the propulsion costs:

$$O = C_f \left( f_s T_s + f_p \sum_p (T_p + T_{p*}) \right) \qquad \text{(Equation 2)}$$

where $T_s$ is the total time spent at sea and and $T_p$, $T_{p*}$ the total times spent at port $p$ within and outside normal work hours respectively and $C_f$ is the fuel unit price. Operational charges also include usage costs:

$$U = C_u \left( T_s + \sum_p (T_p + T_{p*}) \right) \qquad \text{(Equation 3)}$$

with $C_u$ the hourly usage cost (staff) of the ship. Finally, port costs have to be included:

$$P = \sum_p \left( \left\lceil \frac{T_p + T_{p*}}{24} \right\rceil C_p + T_p C_{bp} + T_{p*} C_{bp*} \right) \qquad \text{(Equation 4)}$$

where $\lceil \cdot \rceil$ is a ceiling operator and $C_p$, $C_{bp*}$, $C_{bp*}$, the daily port fee, hourly manutention prices in normal hours and outside normal hours in port $p$ respectively.

$Z$ is the cost of not making the delivery in time (i.e within the horizon $T$). Depending on the aim of the user, $Z$ can also take into account the negative externalities on the environment and society of a grounding ($Z \to -\infty$) or simply the loss for the shipper ($Z = -V - (O + P + U)$).

### 5.2.2 Constraints

The ship's cargo and scheduling have to satisfy some constraints, that we recall here.

1. At any stage, the cargo load $m_j$ cannot exceed the tank capacity $m_{max}$ and must fit with the requirements for safe structural behaviour of the hull ($m_j \geq m_{min}$), as well as with the demand constraints in the next ports to visit ($m_j \geq \sum_{k=j+1}^{N} a_k$). In the following: $m_{min}$ is taken as the minimum of the structural constraint and the economic constraint. The fuel load necessary to carry the ship and its cargo $m_j$ over the distance $l = \sum_{k=j}^{N-1} l_k$ at speed $v$ and load/unload the freight at rate $u_p$ in port $p$ must be subtracted from $m_{max}$: $f_s l + f_p \sum_{p=j+1}^{N} T_p + m_j \leq m_{max}$, where the minimal time spent at port $p$ is the time for (un)loading: $T_p = \frac{|m_{p-1} - m_p|}{u_p}$ (noting that we set $m_0 = 0$).
2. To enter/leave port $p_j$ at time $t$, the water depth must be greater than the ship

draft $r(t)$ plus the safety margin $\alpha r(t)$:

$$\hat{X}_j(t) - (1+\alpha)r(t) > 0. \qquad \text{(Equation 5)}$$

The ship draft is a function of the cargo load as well as the fuel mass $f(t)$ in tanks at the time $t$ of interest. Following Archimedes' principle and the equilibrium of forces in a gravitational field, $r$ can be estimated from the equality between the ship's weight and the water displacement. In a simple approximation (barge ship), we can write:

$$r(t) = \frac{m + f(t) - 0.5m_{max}}{\rho S} + r_{50} \qquad \text{(Equation 6)}$$

where $r_{50}$ is the half laden ship's draft, $S$ the ship's horizontal area, $m_{max}$ its carrying capacity, $\rho$ the water density. The function $f(t)$ is computed by taking into account the fuel consumption rates at sea $f_s$ and at port $f_p$ respectively, the time already spent at sea and at port respectively at $t$, as well as the total fuel load necessary to move the ship from one port to another and (un)load material. Dynamical effects such as the squat effect or the heel due to the wind and the wave responses can reduce the under-keel clearance temporarily. They are not taken into account here beyond the safety margins $\alpha r(t)$ as, again, we consider the still water problem.

3. The ship cannot leave port $p_j$ before the cargo is (un)loaded and must arrive before the horizon $T$ is reached, so:

$$t_{j-1} + \frac{l_{j-1}}{v} + \frac{|m_j - m_{j-1}|}{u_{p_j}} \leq t_j \leq T - \frac{\sum_{k=j}^{N-1} l_k}{v}. \qquad \text{(Equation 7)}$$

### 5.2.3 Sea level predictions

In practice, the time vectors of sea level predictions at a port $p_j$ are decomposed as $\hat{\boldsymbol{X}}_j = \boldsymbol{h}_j + \hat{\boldsymbol{x}}_j$ where $\boldsymbol{h}_j$ is the time series of harmonic tide predictions and $\hat{\boldsymbol{x}}_j$ is the time series of sea-level residual predictions, that is the predictions for the error between observed sea levels $\boldsymbol{X}_j$ and tide predictions $\boldsymbol{h}_j$.

In this paper, we use two sources of information for $\hat{\boldsymbol{x}}_j$:

(A) A best-fit probabilistic/maximum likelihood modelling of $\hat{\boldsymbol{x}}_j$ given joint archives of observations $\boldsymbol{X}_j$ and tide predictions $\boldsymbol{h}_j$ at port $p_j$. Residuals in each port and between successive times are considered independent (a reasonable assumption for British ports not too close, as shown in Rabassa and Beck [2015]) and sampled from the above-mentioned best-fit distributions.

Because in $\hat{X}_j(t) = h_j(t) + \hat{x}_j(t)$, sea level residuals $\hat{x}_j(t)$ are not deterministic predictions but samples from probabilistic predictions, rather than directly optimising the net shipping return, we want to maximise a statistic

$$\mathbb{A}[B(\boldsymbol{d}; \boldsymbol{\Theta}, \{\hat{\boldsymbol{X}}_j\}_{j=1,\ldots,N})]_{\{\hat{\boldsymbol{x}}_j\}_{j=1,\ldots,N}}$$

such as the expected benefit or the worst-case benefit.

(B) Storm-surge ensemble predictions, that provide $M$ time series $\hat{\boldsymbol{x}}_j^m$, $m = 1, \ldots, M$ corresponding to the implementation of $M$ slightly perturbed initial conditions and/or forcing in the hydrological numerical model used to compute storm-surge predictions. These ensemble predictions, synthetically noted $\{\hat{\boldsymbol{x}}_j\}_M$, are typically interpreted by means of standard probabilistic approaches (e.g. Bayesian model averaging [Raftery et al., 2005], non-homogeneous regression [Gneiting et al., 2005], kernel dressing [Roulston and Smith, 2003]) to provide probabilistic predictions of the form $P(\hat{x}_j(t) > c)$ or more generally an estimate of the probability distribution function $p(\hat{x}_j(t))$. From this distribution, residuals in each port can be sampled, similarly to (A) and the optimisation problem consists as well in optimising a statistic $\mathbb{A}_{\{\hat{\boldsymbol{x}}_j\}_{j=1,\ldots,N}}$ of the benefit.

In this work, we use the possibilistic framework developed by the authors in Le Carrer and Ferson [2020] to interpret $\{\hat{\boldsymbol{x}}_j\}_M$ and derive a possibility distribution $\pi(\hat{x}_j(t))$, as will be presented in Sec. 5.3. From there, we need to reformulate the optimisation problem in a fuzzy context, which is introduced in Section 5.4.

## 5.3 A possibilistic framework to interpret ensemble predictions

### 5.3.1 Possibility theory

Possibility theory was developed from fuzzy set theory by Zadeh [1978], Dubois and Prade [2012] as a framework to handle imprecise probabilities. The possibility distribution $\pi : \mathcal{X} \to [0, 1]$ represents a state of knowledge about the state of the system of interest, described by the variable $x \in \mathcal{X}$. When it comes to assess the possibility of observing an event $A = \{x \in S_A\}$, where $S_A \subset \mathcal{X}$, two dual measures are computed. The possibility $\Pi(A) = \sup_{x \in S_A} \pi(x)$ indicates how much $A$ is supported by the evidence at hand, encoded in $\pi$. The necessity $N(A) = 1 - \Pi(\bar{A}) = 1 - \sup_{x \notin S_A} \pi(x)$ indicates how much $A$ is necessary given the evidence at hand (i.e. given the impossibility of observing the complementary event $\bar{A}$). These dual measures satisfy the following axioms (a, b) and conventions (c, d, e) [Cayrac et al., 1994]:

(a) $\Pi(\mathcal{X}) = 1$ and $\Pi(\varnothing) = 0$

(b) $\Pi(A \cup B) = \max\big(\Pi(A), \pi(B)\big)$

(c) $N(A) = 1 \leftrightarrow \Pi(\bar{A}) = 0 :$ $A$ has to happen, it is necessary

(d) $0 < N(A) < 1$ is a tentative acceptance of $A$ to a degree $N(A)$

(e) $\big(\Pi(A) = \Pi(\bar{A}) = 1\big) \leftrightarrow \big(N(A) = N(\bar{A}) = 0\big)$ represents total ignorance: the evidence at hand doesn't allow to conclude whether $A$ is rather true or false.

Probability measure $P$ and possibility measure $\Pi$ are connected through the concept of imprecise probabilities. Several definitions of consistency have been proposed [Delgado and Moral, 1987] and we retain here the view of Dubois et al. [2004]: $P$ and $\Pi$ are consistent if the probability of any possible event $A$ satisfies $P(A) \leq \Pi(A)$. It implies, given the definition of necessity:

$$N(A) \leq P(A) \leq \Pi(A) \qquad \text{(Equation 8)}$$

A possibility distribution $\pi$ is at least as specific as another $\pi'$ when $\pi(x) \leq \pi'(x) \forall x \in \mathcal{X}$. The principle of minimum specificity is the guiding principle in possibility theory [Dubois et al., 2004], aiming at drawing the least possible conservative distributions with a given amount of information.

Finally, we call $\alpha-$cut of the fuzzy number $x$ described by the possibility membership function $\pi$ the set $C_\alpha(x) = \{x \mid \pi(x) \geq \alpha\}$.

### 5.3.2 Possibilistic framework for EPS interpretation

In this section, we summarize the possibilistic framework to interpret EPS and fuse dynamical information introduced and fully developed in Le Carrer and Ferson [2020] by the authors. Figure 5.1 provides a global overview of the framework, explained below.

Considering a dynamical system $\mathcal{S}$ described by a variable $x \in \mathbb{R}$, we assume that we have at our disposal the following elements of information for a prediction at lead time $t$ from $t_0$:

1. A set of $M$ ensemble predictions for lead time $t$, noted $\tilde{\boldsymbol{x}}(t_0+t) = \{\tilde{x}_1(t_0+t), \ldots, \tilde{x}_M(t_0+t)\}$ ;

2. An archive $\mathcal{I}$ containing all the pairs EPS-observations $\{\tilde{\boldsymbol{x}}(t_i+t), x(t_i+t)\}$ for similar lead time $t$ and $N_I$ different starting times $t_i, i = 1, \ldots, N_I$.

3. A time series of (preferably continuous) past observations of $x$, denoted $\mathcal{I}_A$, containing the initial condition (IC) $x(t_0)$ of interest.

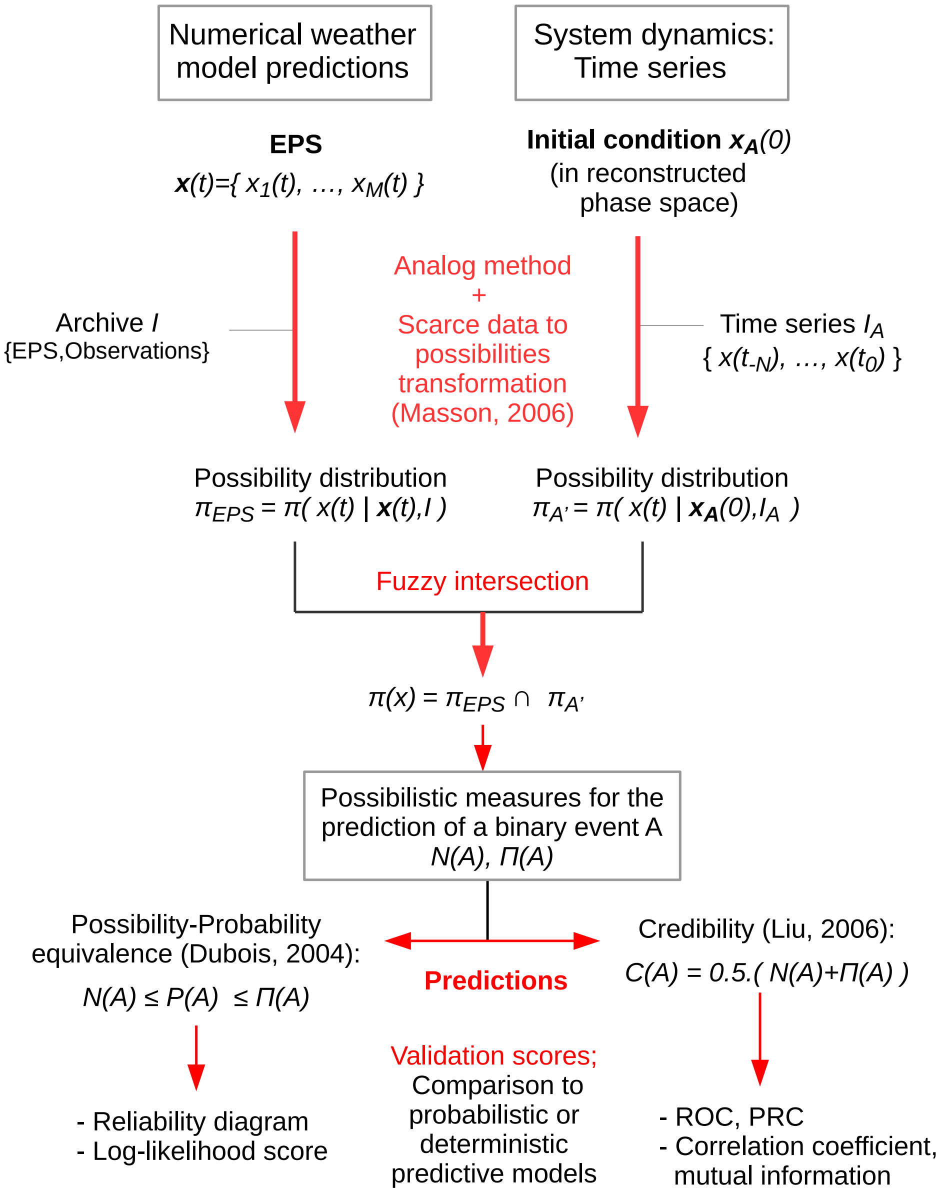

Numerical weather model predictions
System dynamics: Time series
EPS
$\mathbf{x}(t)=\{ x_1(t), \ldots, x_M(t) \}$
Initial condition $\mathbf{x_A}(0)$
(in reconstructed phase space)
Archive $I$
{EPS,Observations}
Analog method
+
Scarce data to possibilities transformation (Masson, 2006)
Time series $I_A$
$\{ x(t_{-N}), \ldots, x(t_0) \}$
Possibility distribution
$\pi_{EPS} = \pi( x(t) \mid \mathbf{x}(t), I )$
Possibility distribution
$\pi_{A'} = \pi( x(t) \mid \mathbf{x_A}(0), I_A )$
Fuzzy intersection
$\pi(x) = \pi_{EPS} \cap \pi_{A'}$
Possibilistic measures for the prediction of a binary event A
$N(A), \Pi(A)$
Possibility-Probability equivalence (Dubois, 2004):
$N(A) \leq P(A) \leq \Pi(A)$
Predictions
Credibility (Liu, 2006):
$C(A) = 0.5.( N(A)+\Pi(A) )$
- Reliability diagram
- Log-likelihood score
Validation scores;
Comparison to probabilistic or deterministic predictive models
- ROC, PRC
- Correlation coefficient, mutual information

**Figure 5.1** General possibilistic framework, from data-driven derivation of possibility distributions to evaluation metrics of predictive skills and the connection with probabilistic predictions.

**Ensemble predictions**

The $x-$axis is first binned in $n$ bins (e.g. uniformly or so that the distribution of the climatology of $x$ over the $x-$axis is uniform). To each bin $b_i$ containing at least an EPS member $\tilde{x}_m$, we associate a possibility distribution $\pi(x|\tilde{x}_m(t_0+t) \in b_i)$. This possibility distribution is constructed using the transformation presented in Masson and Denœux [2006] over the analog set formed by the observations $x(t_j+t)$ associated with ensemble members that fell in bin $b_i$ for predictions of $x(t_j+t)$ for all the ICs $t_j$ contained in the archive $\mathcal{I}$. Each of these possibility distributions represents a partial view on the future system state, given that only the knowledge on one bin filled by one or more EPS member is used. We consequently take the union of these possibility distributions, to compute the final possibility distribution describing the future system state: $\pi_{EPS} = \pi(x(t_0+t)|\boldsymbol{x}(t_0+t), \mathcal{I}) = \cup_i\, \pi(x|\tilde{x}_m(t_0+t) \in b_i)$.

**System dynamics**

$\pi_{EPS}$ encodes the knowledge of the future state of the system gained from the EPS $\tilde{\boldsymbol{x}}(t_0+t)$. However, it lacks information on the IC and local dynamics of the system. We can consequently make it more specific (i.e. less conservative) by combining it to the possibility distribution gained from the analog method applied to time series $\mathcal{I}_A$. Namely, we use the Taken's delay-embedding theorem (or another similarity-based method, e.g. statistical downscaling) to reconstruct the shadow attractor governing the dynamics of the system from vectors $\boldsymbol{x_A}(t) = \big(x(t), x(t-\tau), \ldots, x(t-(m+1)\tau)\big)$ where the embedding dimension $m$ and the time-delay $\tau$ are determined by means e.g. of the simplex method [Sugihara et al., 2012]. From there, for each prediction of interest, we locate the IC $\boldsymbol{x_A}(t_0)$ in the reconstructed shadow attractor, find the $n_A$ closest neighbors in terms of Euclidean distance and follow their trajectory up to lead time $t$, which provides us with a set of $n_A$ analogs of the future state of the atmosphere. Again, we use the transformation developed in Masson and Denœux [2006] to derive the corresponding possibility distribution $\pi_{A'}$.

We finally combine $\pi_{EPS}$ and $\pi_{DYN}$ by taking their fuzzy intersection $\pi_{COMB} = \pi_{EPS} \cap \pi_{DYN}$, given by the min-envelope $\pi_{COMB}(x) = \min\big(\pi_{EPS}(x), \pi_{DYN}(x)\big)$ [Zadeh, 1975]. This allows to reduce their respective over-conservatism due to the incomplete knowledge encoded in each one.

Note that the data-to-possibility transformation that we use is designed to account for the uncertainty due to limited datasets (here, the archives $\mathcal{I}$ and $\mathcal{I}_A$).

**From possibility distribution to prediction**

From the resulting predictive possibility distribution $\pi_{COMB}$ (or $\pi$ for short), and an event of interest $A$, one can extract the possibility and necessity measures $\Pi(A)$ and $N(A)$ as described in Sec. 5.3.1. These can be used to make an imprecise probabilistic prediction of $A$ by means of Equation 8. Or, we can deduce from them the credibility indicator $C(A) = \frac{N(A)+\Pi(A)}{2}$, aggregating both measures [Liu, 2006]. Associated to relative operating characteristics or precision-recall curves, the credibility allows to make deterministic predictions (yes/no). Associated to actual frequencies of observations, they can be used to propagate frequentist probabilities in subsequent applications [Le Carrer and Green, 2020].

## 5.4 Fuzzy-constrained optimisation problem

In Section 5.2, we described the maritime shipping optimisation problem at hand. Namely, given a set of deterministic tide predictions $h(t)$ and sea level residual predictions $\hat{x}_j(t), t \in [t_0, t_0+T]$ for the entrance channel of each port of call $p_j, j = 1, \dots, N$ as well as a range of *a priori* fixed journey parameters $\boldsymbol{\Theta}$, we want to find at $t_0$ the decision $\boldsymbol{d} = \big(m_1, t_1, \dots, m_{N-1}, t_{N-1}\big)$ that maximises a statistic $\mathbb{A}_{\{\hat{\boldsymbol{x}}_j\}_{j=1,\dots,N}}$ of the shipping return $B(\boldsymbol{d}; \boldsymbol{\Theta}, \{\hat{\boldsymbol{X}}_j\}_{j=1,\dots,N})$ where $\hat{X}_j(t) = h_j(t) + \hat{x}_j(t)$.

We now describe an approach to solve this problem in the case when $\hat{x}_j(t)$ is given through ensemble predictions interpreted in a possibilistic way, as described in Section 5.3. The subsequent possibilistic nature of the time series $\hat{X}_j(t)$ (given that $h_j(t)$ is deterministic and acts as an additive constant), impacts the problem at two levels: (i) verification of the constraints (Equation 5) in each port $p_j$, that we generically note $g_j(\mathbf{d}, \hat{\boldsymbol{X}}_j) > 0$; and (ii) the computation of the net benefit $B(\boldsymbol{d}; \boldsymbol{\Theta}, \{\hat{\boldsymbol{X}}_j\}_{j=1,\dots,N})$.

To address the optimisation problem in a framework adapted to the possibilistic nature of some of its parameters, we follow the procedure described in Hose et al. [2018, 2019], that we recall here:

1. **Confluence of constraints** For a decision $\boldsymbol{d}$, the joint necessity of verifying all constraints $g_j, j = 1, \dots, N$ is provided by the intersection of their marginal necessities:

$$\sigma(\boldsymbol{d}) = \min_{j=1,\dots,N} N\big(g_j(\boldsymbol{d}, \hat{\boldsymbol{X}}_j) \succeq 0\big) \qquad \text{(Equation 9)}$$

   where $\succeq$ indicates 'above or equal' in a fuzzy context.

2. **Maximum compliance with constraints** A global optimiser (in our case particle swarm optimisation, as recommended by Hose et al. [2018]) is used to find the highest achievable value of the combined necessities:

$$\sigma^* = \max_{\boldsymbol{d} \in \mathcal{D}} \sigma(\boldsymbol{d}) \qquad \text{(Equation 10)}$$

   where $\mathcal{D}$ defines the search space for decisions $\boldsymbol{d}$, defined by other crisp constraints or upper/lower bounds on elements of $\boldsymbol{d}$.

3. **Feasible set** The feasible set $\mathcal{R}$ of decisions that achieve maximum robustness

is consequently:

$$\mathcal{R} = \{\boldsymbol{d} \in \mathcal{D} \mid \sigma(\boldsymbol{d}) = \sigma^*\} \qquad \text{(Equation 11)}$$

Note that the feasible set can also be chosen wider by including all decision variables that attain a minimal necessity $\beta$:

$$\mathcal{R} = \{\boldsymbol{d} \in \mathcal{D} \mid \sigma(\boldsymbol{d}) \geq \beta\} \qquad \text{(Equation 12)}$$

This means that the probability that the constraints are jointly fulfilled is at least $\beta$ (cf. Equation 8).

4. **Minimisation of the maximum possible error** Following Hose et al. [2019], Jamison and Lodwick [1999], worst-case optimality is achieved by finding:

$$\boldsymbol{d}^* = \arg\min_{\boldsymbol{d}\in\mathcal{R}} \max_{\boldsymbol{X}\in C_{1-\sigma^*}(\tilde{\boldsymbol{X}})} [B(\boldsymbol{d};\boldsymbol{\Theta},\boldsymbol{X}) - \inf_{\boldsymbol{y}\in\mathcal{R}} B(\boldsymbol{y};\boldsymbol{\Theta},\boldsymbol{X})] \qquad \text{(Equation 13)}$$

where $\tilde{\boldsymbol{X}}$ represent the joint fuzzy variables $\{\boldsymbol{X}_j\}_{j=1,\ldots,N}$ and $\boldsymbol{X}$ a sample from $\tilde{\boldsymbol{X}}$. These realisations of the fuzzy sea levels can be drawn from their respective $1-\sigma^*$ cuts only since realisations with lower degrees of memberships are not guaranteed to verify the constraints.

Because we deal with high-dimensional time series of fuzzy parameters (the sea levels in each port $j$), directly applying this formula is not computationally workable. We consequently intertwine the overall procedure described above with the possibilities of Particle Swarm Optimisation (PSO) to estimate $\boldsymbol{d}^*$.

Namely, we use a two-level PSO, whose first level drive particles in the search space $\mathcal{D}$ towards the areas of maximal compliance $\sigma(\boldsymbol{d})$. The secondary level consists, at each step of the algorithm, in computing, for the particles reaching or exceeding their personal best in terms of maximal compliance, the statistic $\mathbb{A}$ on the net benefit from samples of the corresponding $1-\sigma^*$ cut:

$$\mathbb{A}[B(\boldsymbol{d};\boldsymbol{\Theta},\boldsymbol{X})]_{\boldsymbol{X}\in C_{1-\sigma^*}(\tilde{\boldsymbol{X}})} \qquad \text{(Equation 14)}$$

where $\sigma^*$ is the particle's best compliance. This statistics is stored for each particle as their second-level personal best. This allows the algorithm to progress towards areas of maximal compliance and within them, to favour positions maximising the statistic (Equation 14) on the benefit.

Overall, the algorithm, with the limits of a heuristic in terms of guarantees, converges towards the decision $\boldsymbol{d}$ that maximises the compliance to constraints and within this maximum, optimises the statistic $\mathbb{A}$ on $B$. On this basis, it is also possible to find with a similar procedure the decision that optimises $\mathbb{A}[B]$ and only ensures that $\sigma(\boldsymbol{d}) \geq \beta$.

To compute the statistics $\mathbb{A}$ on the benefit $B$, for each candidate decision $\boldsymbol{d}$, we retrieve the $1 - \sigma$ cuts (where $\sigma$ is the global level of constraint compliance for $\boldsymbol{d}$) of the corresponding sea levels in each port. The idea behind this choice is that realizations of the fuzzy sea levels with lower memberships than $1 - \sigma$ are not guaranteed to fulfill the fuzzy-valued set of constraints and consequently sea levels need only to be drawn from their respective $1 - \sigma$ cuts. Following the idea of the transformation method [Hanss, 2005], we discretise them into $n_b = 25$ bins of uniform width. Then, we compute the benefit for all possible combination of the discretised sea levels. The number of calculations amounts to $n_b^N$, hence $n_b$ is clearly a trade-off between precision and computational cost. Finally, from this set of 'sampled' benefits, we evaluate $\mathbb{A}$.

## 5.5 Experimentation and Results

Although the model is developed for a $N$-port maritime inventory routing problem, this case study addresses the case $N = 2$ ports (due the the limited database of storm surge ensemble predictions at our disposal). Namely, we consider a farm cooperative that owns a small-size bulk carrier and carries regularly malting barley freights from Liverpool to Southampton. Table 5.1 gathers the contextual parameters regarding the shipping problem, including ship characteristics, freight and port management, generic constraints about acceptable under-keel clearance, latest arrival time and cargo load, demand constraints in delivery ports. The minimal depth guaranteed in each port is assumed to be 12 (Liverpool) and 7 m (Southampton). We have at our disposal one year (2017) of tide predictions and associated observations, sampled every $\Delta t = 15$ minutes in both ports, provided by the British National Tidal and Sea Level Facility (British Oceanographic Data Centre, Environment Agency). In addition, we have access to 6 months (July-December 2017) of the storm surge ensemble predictions produced by the British Oceanographic Data Centre [Flowerdew et al., 2010, 2013]. Ensembles of $M = 23$ members are produced by running the CS3X storm surge model [Flather, 2000] with $M$ perturbed forcing conditions, provided by the ensemble members from the Met Office Global and Regional Ensemble Prediction System [Bowler et al., 2008]. They are sampled with the same time step $\Delta t$.

### 5.5.1 Value of EPS possibility-based predictions

Before analysing the value of the possibilistic interpretation of EPSs in the optimisation problem, we briefly compare the predictive performances of our possibilistic

**Table 5.1** Model parameters

| Type | Param. | Description | Value | Unit |
|---|---|---|---|---|
| Journey | $l$ | Mean distance between departure and arrival ports | 440 | Nautical miles |
| | $\rho$ | Mean sea water density | 1,250 | Kilogram per cubic meter |
| Ship design | $v$ | Mean operational sailing speed | 13 | Knot |
| | $S$ | Ship horizontal surface | 25 × 130 | Meter×Meter |
| | $m_{min}$ | Minimum cargo load (ballast) | 3,000 | Metric ton |
| | $m_{max}$ | Deadweight tonnage (carrying capacity) | 25,000 | Metric ton |
| | $r_{50}$ | Half-laden ship draft | 8 | Meter |
| | $f_s$ | Fuel consumption rate of the laden ship at sea | 11 | Ton per day |
| | $f_p$ | Fuel consumption rate of the ship at port | 2 | Ton per day |
| Monetary | $C_f$ | Fuel cost | 387 | US$ per ton |
| | $C_u$ | Other operational costs (staff, maintenance) | 2,500 | US$ per day |
| | $C_c$ | Average bulk cargo value | 195.6 | US$ per ton |
| | $C_{bp*}$ | Berthing and loading/ unloading operation cost within normal opening times | $\{1, 2391, 486\}$ | US$ per hour |
| | $C_{bp*}$ | Berthing and loading/ unloading operation cost outside of normal opening times | $\{1, 5481, 858\}$ | US$ per hour |
| | $C_p$ | Daily port fee | $\{1, 1151, 363\}$ | US$ per day |
| Port | $u_p$ | Bulk material (un)loading rate | $\{1, 2001, 000\}$ | Ton per hour |
| | | Normal port opening time | $[7:00, 19:00]$ in all ports | - |
| | $\alpha$ | Minimum allowed under-keel clearance to navigate in port still waters | 10% static draft | - |
| Forecast | $\Delta t$ | Sea level forecast time step | 15 | Minute |
| | $T$ | Horizon of the sea level predictions | 6 | Day |
| Industrial | $a_j$ | Minimal delivery in port $j > 1$ | 4,000 | Metric ton |

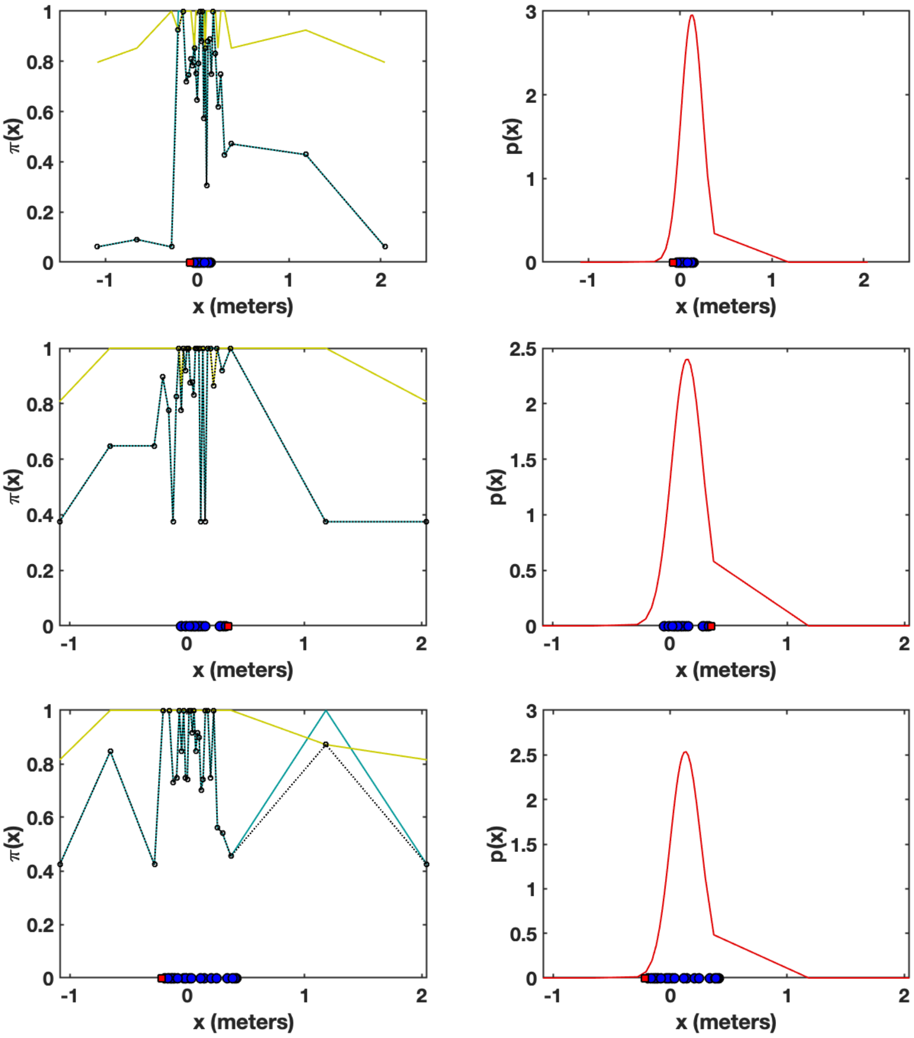


**Figure 5.2** For a given EPS (blue dots) and the actual observation (red square) in Liverpool port, we represent the predictive possibility distributions (EPS in solid blue line, DYN in solid yellow line and COMB in dotted black line) as well as the probabilistic prediction (right, in red line). From top to bottom, lead times are 2, 3 and 4 days.

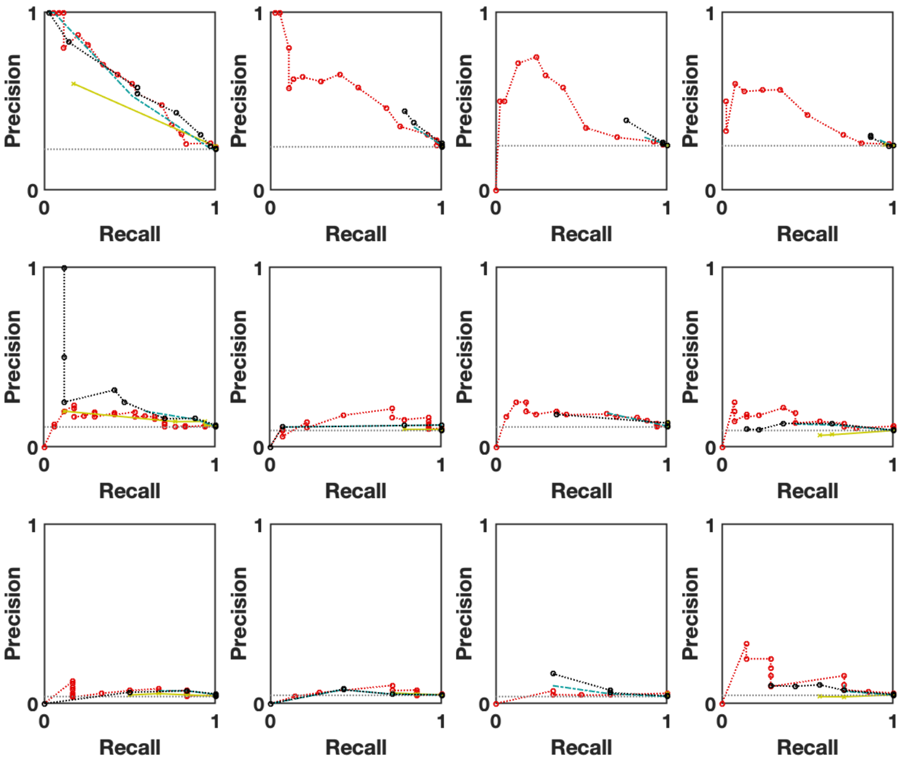


**Figure 5.3** PRC plots for the prediction of the three events A, B, C (top to bottom) over 52 days in Southampton port.

framework with respect to a standard Gaussian ensemble dressing. This probabilistic interpretation, assuming that the members are exchangeable, consists in fitting a parametric probability distribution around each linearly corrected ensemble member and summing them all to provide the global density function. It reads [Roulston and Smith, 2003]:

$$p(x|\tilde{\boldsymbol{x}})_\theta = \frac{1}{M}\sum_{i=1}^{M}\mathcal{N}(a\tilde{x}_i + \omega, \sigma^2) \qquad \text{(Equation 15)}$$

where $\mathcal{N}(\mu, v)$ is the normal distribution of mean $\mu$ 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 $100$ pairs {EPS,observations} for each lead time of interest $t = \{1, 2, 3, 4\}$ days (starting times are separated of two days). To account for the variability of results from one set to the other, we repeat the optimisation procedure $10$ times on different subsets of size $90$ of the whole training set. We then use the resulting $10$ sets of parameters to compute the performance metrics relative to the probabilistic approach on a test set of size $52$. Finally, we take the average of these

10 scores, that we report on the graphs as representative of the performances of the probabilistic approach. Figure 5.2 illustrates the resulting probabilistic and possibilistic distributions for various lead times in Liverpool port.

We now present the precision-recall curves (PRC) associated to the predictions of the events $A = \{x(t) \leq 0\}$ cm, $B = \{x(t) \geq 25\}$ cm and $C = \{x(t) \geq 50\}$ cm for the port of Southampton, where we omit the reference to the IC $t_0$ and simply consider the lead time $t$. PRC are used to estimate the ability of a predictive model to discriminate between event and non-event in the case of imbalanced data sets. The precision (rate of correctly predicted $A$ over all $A$ predicted) is plotted as a function of the recall rate (a.k.a. hit rate, fraction of correctly predicted $A$ over all $A$ observed). In both probabilistic and possibilistic cases, we use increasing thresholds $p_l \in [0, 1]$ for making the decision ($A$ predicted if $P(A) \geq p_l$ (resp. $C(A) \geq p_l$) 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.

Figure 5.3 presents the corresponding results for lead times 1 to 4 days in Southampton port. Conclusions hold for Liverpool sea level time series at hand. The figure reports the PRC associated to the probabilistic approach defined above, as well as to the possibilistic approach based on the EPS only ($\pi_{EPS}$) and the EPS combined to the dynamical information ($\pi_{EPS \cap DYN}$). Due to the small size of the test set ($52$ samples for each lead time of interest), these results are qualitative mostly. Yet, we can note that:

- Overall, the probabilistic approach tends to perform equivalently or better in terms of discrimination. Performances of the possibilistic and probabilistic approaches are globally close, however in some cases the probabilistic approach allows to reach higher levels of precision (at the cost of a lower recall) while the possibilistic does not.
- Generally, $\pi_{EPS}$ performs better than $\pi_{DYN}$ and yet it can be slightly improved by their combination $\pi_{COMB}$.
- Prediction skill (here discrimination), for such a small archive, tends to decrease with more extreme events for both probabilistic and possibilistic methodologies.

These observations are due to the fact that the possibilistic approach is similarity-based, for both EPS and dynamical sources of information, hence it performs all the more that the size of the archive is large (to a certain extent when it comes to extreme events, as discussed in our previous work [Le Carrer, 2021]). Here the size of the archive is of $100$ elements, which is very small for analog-based methods. The possibilistic approach consequently classifies most of the predictions in the ignorance

| Lead time (days) | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| $\pi_{EPS}$ | 0.40 | 0.49 | 0.64 | 0.81 |
| $\pi_{COMB}$ | 0.01 | 0.11 | 0.12 | 0.16 |

**Table 5.2** Proportion of possibilistic predictions falling in the ignorance area for the event $B = \{x(t) \geq 25\}$ cm in Southampton port.

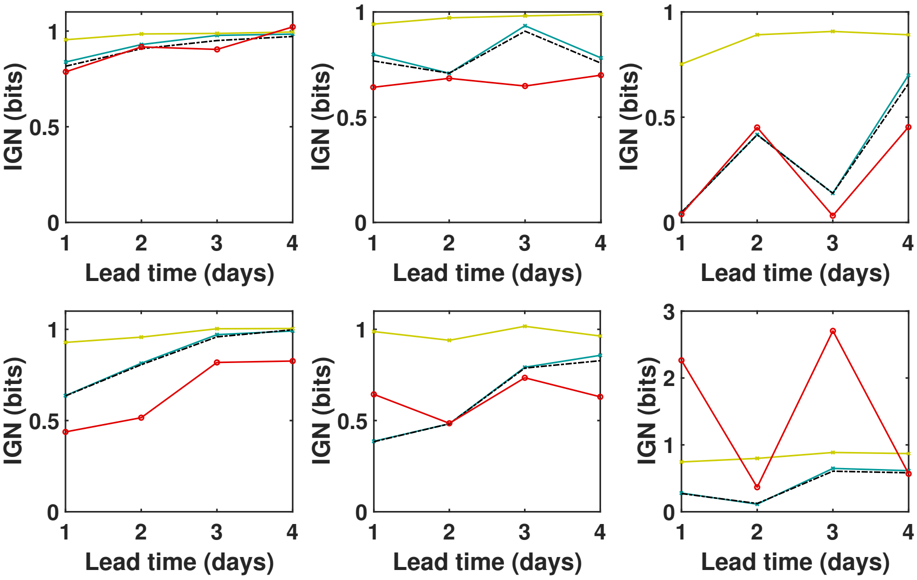


**Figure 5.4** Ignorance score of the possibilistic predictive methods EPS, DYN and COMB and of the probabilistic prediction in both Southampton (top) and Liverpool (bottom).

area (see ratios in Table 5.2), which leads to high recall with very low discrimination skills. However, especially for medium and large recalls, the possibilistic approach performs quite similarly to the probabilistic one.

Another way of comparing predictive performances of probabilistic predictions is to use the ignorance score. The score 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:

$$S_N(G) = \frac{1}{N}\sum_{i=1}^{N} -\log_2 G(O_i) \qquad \text{(Equation 16)}$$

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. We use the credibility $C(O_i)$ in place of $G$ in the possibilistic case. The larger the ignorance, the less interesting the prediction methodology.

Figure 5.4 presents the metric for lead time 1 to 4 days, for both ports and the three events of interest. These plots confirm what we observed with the PRC plots: 1) Globally, the probabilistic method performs similarly or better than our possibilistic approaches, apart from the extreme event C case in Liverpool; 2) Using the combination with the dynamical source of information does add little information to the EPS-based prediction.

The next section will allow us to see if used in an optimisation problem the possibilistic predictions can yet lead to advantages (e.g. worst-case guarantees, robustness) with respect to the probabilistic approach.

### 5.5.2 Experience 1

As detailed in Section 5.2.3, for each port $j$, in addition to the deterministic tide predictions $\boldsymbol{h}_j$, we have at hand two sources of information:

(A) A best-fit probabilistic/maximum likelihood modelling of the residual time series $\hat{\boldsymbol{x}}_j$, given joint archives of observations and tide predictions;

(B) The ensemble predictions for residuals, $\{\hat{\boldsymbol{x}}_j\}_M$.

This leads to five different methodologies, allowing to assess the respective value of each source of information when it comes to the shipping optimisation problem:

(a) A probabilistic optimisation of the decision vector $\boldsymbol{d}$ based on the best-fit probabilistic modelling of the residuals $\hat{\boldsymbol{x}}_j$. It consists in finding the decision $\boldsymbol{d}$ that optimises a statistic $\mathbb{A}[]_{\{\hat{\boldsymbol{x}}_j\}_{j=1,\ldots,N}}$ of the net shipping return $B(\boldsymbol{d}; \boldsymbol{\Theta}, \{\hat{\boldsymbol{X}}_j\}_{j=1,\ldots,N})$. For each point $\boldsymbol{d}$ in the search space, the statistic $\mathbb{A}[]_{\{\hat{\boldsymbol{x}}_j\}_{j=1,\ldots,N}}$ is computed by means of Monte-Carlo sampling, where $N_{MC}$ residuals are sampled independently for each time step and between ports, from their respective best-fit distributions. The corresponding net benefits are computed and the resulting statistic is estimated from them. A particle swarm optimisation (PSO) algorithm is used to to find the optimal decision $\boldsymbol{d}$. This method was developed and described in Le Carrer et al. [2020] and is refered to in the Figures as *Prob. tides*.

(b) A similar procedure of probabilistic optimisation of $\boldsymbol{d}$, this time based on the probabilistic modelling of the residuals $\hat{\boldsymbol{x}}_j$ derived from the EPS. We refer to it as *Prob. EPS*.

(c) The fuzzy-constrained optimisation procedure developed in Section 5.4, that uses the possibilistic interpretation of the ensemble predictions $\{\hat{\boldsymbol{x}}_j\}_M$. The algorithm, with the limits of a heuristic in terms of guarantees, converges towards the decision $\boldsymbol{d}$ that maximises the compliance to constraints $\sigma$ (or ensure that $\sigma$ is above a given level $\beta$) and within this maximum, optimises the statistic $\mathbb{A}$ on the net shipping return $B$. We refer to this method as *Poss. EPS.*

(d) A deterministic optimisation using the tide predictions only, without considering residuals. It provides the decision $\boldsymbol{d}$ that optimises the net benefit $B(\boldsymbol{d}; \boldsymbol{\Theta}, \{\boldsymbol{h}_j\}_{j=1,\ldots,N})$. We use again a PSO algorithm for global optimisation.

(e) The same deterministic optimisation based on tide predictions only, yet incorporating systematically a rule-of-thumb safety margin of $s$ m, as most often done in practice. In other words sea level predictions read $\boldsymbol{h}_j - s$. We refer to it as *Det. SM=* $s$.

The last two methodologies are here to 1) help identify the limitation of using tide predictions only in terms of shipping optimisation, and 2) assess the potential added value of rather computationally intensive probabilistic or possibilistic approaches over the simple rule-of-thumb safety margin approach used on the field.

We sample randomly 30 days between $01/07/2017$ and $31/12/2017$. On each of these days, we get at $6:30$ GMT the time series over the next $7$ days for tide predictions and ensembles of residual predictions. We compute the optimal decision $\boldsymbol{d}$ provided by each methodology, for the shipping problem developed in Section 5.2 and the corresponding actual benefit $B(\boldsymbol{d}; \Theta, \boldsymbol{X}_j, j = 1 \ldots N)$. We compare these result to the respective expected benefits $B(\boldsymbol{d}; \Theta, \hat{\boldsymbol{X}}_j, j = 1 \ldots N)$ and to the optimal benefit that an user would get if he knew perfectly the future sea levels, that is the benefit corresponding to the decision maximising $B(\boldsymbol{d}; \Theta, \hat{\boldsymbol{X}}_j, j = 1 \ldots N)$.

We consider two statistics $\mathbb{A}$: the expected benefit and the worst case, following previous work in [Le Carrer et al., 2018] and their respective advantages (the expected benefit performs better in average, however the worst case is expected to be more robust and avoids rare but potentially catastrophic situations).

Figure 5.5 shows the median and standard deviation of the relative difference between actual benefit of a journey given a predictive methodology and the benefit obtained with perfect knowledge of future sea levels. In other words, we assess the relative loss (of benefit) induced by imperfect predictions.

We first observe that the purely deterministic option (SM=0) leads to the average best actual benefits, very close to perfect-information based ones. However the method occasionally leads to dramatic decisions. Using a safety margin as low as $25$ cm allows to remove this limitation. Naturally, the larger the safety margin, the lower the average subsequent benefit. As noted in [Le Carrer et al., 2020], the probabilistic modeling of residuals leads to actual benefits similar in distribution to those provided by a fixed deterministic safety margin, in this problem of $0.75$ m. One could consequently argue that it may be cheaper to set a fixed margin instead of running more complex and costly

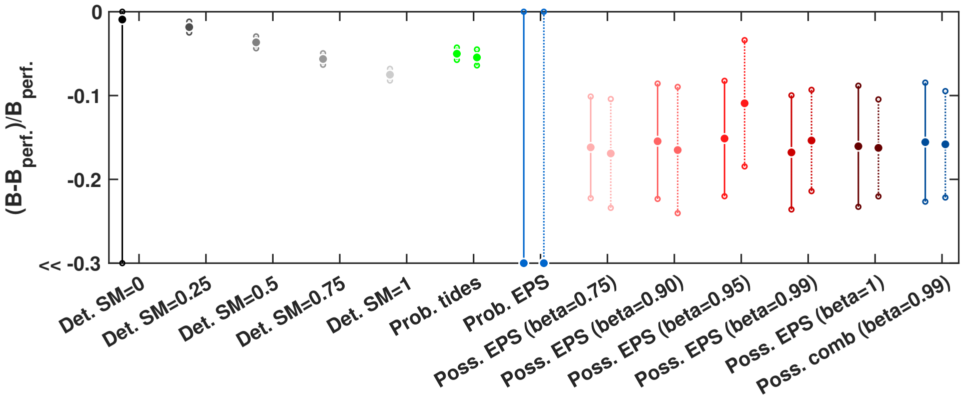


**Figure 5.5** Statistics (median plus/minus standard deviation over 30 journeys) of the relative difference between actual benefit of a journey given a predictive methodology and the benefit obtained with perfect knowledge of future sea levels. For readability, we only show the range $[-0.3, 0]$ however the lower bound of the 'Det. SM=0' method is $-2.1$, just like the average of 'Prob. EPS'. Upper bound are 0 by definition of the perfect decision hence we truncate the upper branch of each plot when necessary. For the methodologies taking into account the stochasticity of tide residuals we report results when the expected benefit (solid line, left) or the worst-case benefit (dotted line, right) are used as objective function.

optimisation algorithms doubled with Monte-Carlo sampling. This is true *a posteriori*, once simulations are run for a given problem (set of ports) and that we can estimate the optimal safety margin. however the safety-margin is problem dependent (see different results on our slightly different case study [Le Carrer et al., 2020]) and less robust than the probability-based method to extreme variations of the residuals (compared to their distributions, initially considered as stationary), as shown in our previous work [Le Carrer et al., 2020]. **The probabilistic approach remains consequently more attractive and optimal in practice.**

When it comes to the ensemble forecasts, treating them in a probabilistic way leads to almost sure losses (i.e. very poor decisions) while the possibilistic approach, although in all its forms is less performing and more variable than (d), (e) and (a), maintains a reasonable benefit compared to the perfect situation. On our test set, we cannot really detect any impact of the confidence level $\beta$ in the constraints to be satisfied, nor of the addition of dynamical information. The latter is in agreement with our previous observations on Figures 5.3 and 5.4.

Using as operator $\mathbb{A}$ the expected benefit or the worst case has very little impact on the probabilistic approach apart from lowering slightly the average benefit and limiting variations in the second case. When it comes to the possibilistic approaches, a small impact can be found on both mean and variance but without clear trend according to $\beta$.

Figure 5.5 shows how robust is a methodology, namely: what is the distribution

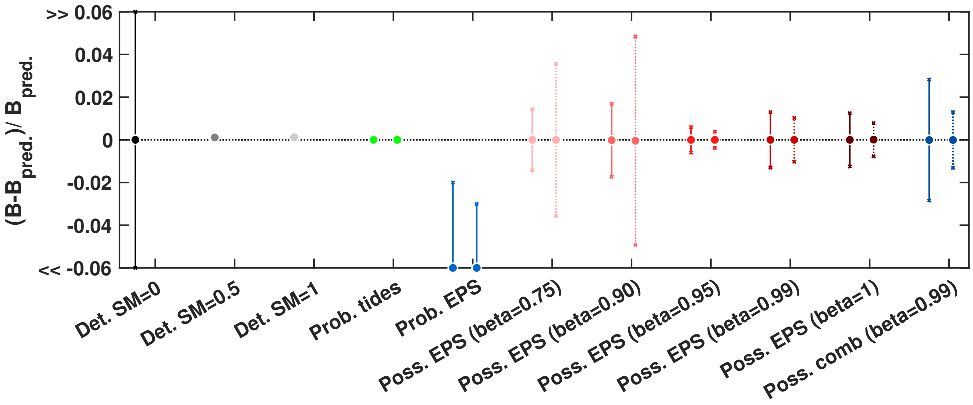


**Figure 5.6** Statistics (median plus/minus standard deviation over 30 journeys) of the relative difference between actual benefit of a journey given a predictive methodology and the benefit predicted by the same methodology. For readability, we only show the range $[-0.06, 0.06]$ however the lower bound of the 'Det. SM=0' method is $-2$, just like the average of 'Prob. EPS'. For the methodologies taking into account the stochasticity of tide residuals we report results when the expected benefit (solid line, left) or the worst-case benefit (dotted line, right) are used as objective function.

of the difference between the benefit *a priori* predicted from the methodology and the *a posteriori*, actual benefit. We note that all deterministic approaches with safety margins are very stable. The relative difference, if existing, is slightly in favor of the actual benefit. Similarly, the probabilistic modelling of residuals leads to unsurprising actual benefits. These results are in agreement with our previous developments in Le Carrer et al. [2020]. As suggested in Figure 5.5, using the EPS in a probabilistic manner leads to surprising and dramatic outcomes most often. Using the worst-case $\mathbb{A}$ statistics allows occasionally to get rather acceptable, decisions, however in general trusting its outcomes leads to serious loss compared to the expected benefit. The possibilistic approach shows in average unsurprising benefits, however the variance is not negligible compared to the former methodologies. We note the impact of the operator $\mathbb{A}$: using the expected benefit as objective function (rather than the worst-case benefit) tends to lead to more predictable actual benefits for $\beta \geq 0.95$ and the reverse is observed for smaller $\beta$. In the latter case, the variance of the difference between predicted and actual benefit is much bigger which explains why a conservative worst-case objective function allows to stabilize a bit more results. **It also shows that sacrifying the confidence level on constraint satisfaction ($\beta < 1$) has a cost: the benefit predictions are (on a one-case basis) less robust, although in average the actual benefit is little modified with varying $\beta$.**

Overall, the possibilistic methodology as implemented in this work is more conservative than using a simple and static probabilistic modelling of sea level residuals. The level of confidence $\beta$ does impact little the average net benefit of induced decisions. However it does impact the robustness of these decisions. Yet, because of the conservatism of the method, decisions are never dramatic as they can be in the purely deterministic approach or with the probabilistic treatment of the EPS information. Our study is based on a very small set of simulations, hence cautious extrapolations. Nevertheless we can note a trend according to which $\beta \approx 0.95$ would be an optimal choice: average actual benefits are the higher, sometimes better than the PROB approach, while robustness of predicted benefits is the best. This is noted with both tested statistics $\mathbb{A}$, however it is all the more true with when we use the expected benefit as objective function. For lower $\beta$, the average benefit decreases as well as the robustness of decisions. For lower $\beta$, both the robustness and average benefit are slightly lower.

## 5.6 Conclusion

Following promising results presented in Le Carrer and Green [2020], Le Carrer [2021] and Le Carrer and Ferson [2020], we tested in this paper an application of the possibilistic interpretation of weather forecasts, namely on sea level residuals and their use in the optimisation of ship routing decisions. To that purpose, we adapted a methodology developed for possibilistic optimisation [Hose et al., 2018] to the problem of ship routing under possibilistic sea levels.

The results presented here confirm, when it comes to the predictive performance of the possibilistic interpretation of EPS, that the size of the archive (EPS-observation) at hand matters for such a similarity-based approach. For very small archives, a probabilistic interpretation will generally work equally well or better, at the exception of extreme events. This finding confirms what we found in the above-mentioned previous works, although it is not as marked in this study due to the very small test dataset at hand ($100$ versus $40.10^3$ in Le Carrer [2021]).

When it comes to the application however, we found that, when it comes to using EPS as source of information on sea levels, only a possibilistic interpretation led to reasonable decisions w.r.t. ship routing. The probabilistic approach, due to its lack of conservatism (contrary to the frequent classification of predictions in the 'ignorance' area in the possibilistic case), tends to lead to biased predictions and decisions that are consequently sure fails.

Beyond that and to answer our initial question, this small experiment shows that the most optimal way of taking ship routing decisions among our propositions is a simple optimisation based on static, historical best-fit modelling of sea level residuals. This equals to adding a rule-of-the-thumb safety margin to the ship's draft, without the need to find out this problem-dependent, and less robust to non-stationarity of sea level distributions, safety margin. Our possibilistic approach, in these particular conditions of extremely limited archive, is too conservative (ignorance prevails) to compete with the probabilistic modelling above-mentioned. However, the analysis of the effect of parameter $\beta$, namely the confidence level in constraint satisfaction, suggests that a careful choice (here $\beta = 0.95$) already allows to achieve decisions sometimes challenging those given by the probabilistic method.

Repeating the experiment with a larger archive of EPS predictions (typically 2-3 years instead of 6 months, cf. experiments presented in Le Carrer [2021]) would consequently be an interesting future work. We expect an improvement of the average possibility-based results, which means that those with well-chosen $\beta$ could challenge more seriously the probabilistic results.

Beyond that, another interesting work would be to compare the effect of the number of ports in the decision-making problem on these comparative performances. Taking into account several ports at a time increases the conservatism of the possibility distribution accounting for the satisfaction of all constraints Hose [2020]. This could be a serious limitation with respect to the probabilistic approach who suffers less of this asymptotic behaviour.

## Bibliography

Agostinho Agra, Marielle Christiansen, Alexandrino Delgado, and Lars Magnus Hvattum. A maritime inventory routing problem with stochastic sailing and port times. *Computers and Operations Research*, 61:18–30, 2015.

Amir Azaron and Farhad Kianfar. Dynamic shortest path in stochastic dynamic networks: Ship routing problem. *European Journal of Operational Research*, 144(1): 138–156, 2003.

Neill E Bowler, Alberto Arribas, Kenneth R Mylne, Kelvyn B Robertson, and Sarah E Beare. The mogreps short-range ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography*, 134(632):703–722, 2008.

Jochen Bröcker and Leonard A. Smith. From ensemble forecasts to predictive distribution functions. *Tellus A: Dynamic Meteorology and Oceanography*, 60(4):663–678, 2008. doi: 10.1111/j.1600-0870.2007.00333.x.

Roberto Buizza. Ensemble Forecasting and the Need for Calibration. In *Statistical Postprocessing of Ensemble Forecasts*, pages 15–48. Elsevier, 2018. ISBN 978-0-12-812372-0. doi: 10.1016/B978-0-12-812372-0.00002-9.

D. Cayrac, D. Dubois, M. Haziza, and H. Prade. Possibility theory in "Fault mode effect analyses". A satellite fault diagnosis application. In *Proceedings of 1994 IEEE 3rd International Fuzzy Systems Conference*, pages 1176–1181. IEEE, 1994.

Tzung-Nan Chuang, Chia-Tzu Lin, Jung-Yuan Kung, and Ming-Da Lin. Planning the route of container ships: A fuzzy genetic approach. *Expert Systems with Applications*, 37(4):2948–2956, 2010.

Hans de Vries. Probability forecasts for water levels at the coast of the netherlands. *Marine geodesy*, 32(2):100–107, 2009.

S De Zolt, P Lionello, A Nuhu, and A Tomasin. The disastrous storm of 4 november 1966 on italy. 2006.

M. Delgado and S. Moral. On the concept of possibility-probability consistency. *Fuzzy Sets and Systems*, 21(3):311–318, 1987. ISSN 0165-0114. doi: https://doi.org/10.1016/0165-0114(87)90132-1.

Tom Di Liberto, Brian A Colle, Nickitas Georgas, Alan F Blumberg, and Arthur A Taylor. Verification of a multimodel storm surge ensemble around new york city and long island for the cool season. *Weather and forecasting*, 26(6):922–939, 2011.

Didier Dubois and Henri Prade. *Possibility theory: an approach to computerized processing of uncertainty*. Springer Science and Business Media, 2012.

Didier Dubois, Laurent Foulloy, Gilles Mauris, and Henri Prade. Probability-Possibility Transformations, Triangular Fuzzy Sets, and Probabilistic Inequalities. *Reliable computing*, 10(4):273–297, 2004. doi: 10.1023/B:REOM.0000032115.22510.b5.

Roger A Flather. Existing operational oceanography. *Coastal Engineering*, 41(1-3): 13–40, 2000.

Jonathan Flowerdew, Kevin Horsburgh, Chris Wilson, and Ken Mylne. Development and evaluation of an ensemble forecasting system for coastal storm surges. *Quarterly Journal of the Royal Meteorological Society*, 136(651):1444–1456, 2010.

Jonathan Flowerdew, Ken Mylne, Caroline Jones, and Helen Titley. Extending the forecast range of the uk storm surge ensemble. *Quarterly Journal of the Royal Meteorological Society*, 139(670):184–197, 2013.

Wiesław Galor. Determination of dynamic under keel clearance of maneuvering ship. *Journal of KONBiN*, 8(1):53–60, 2008.

Herman Gerritsen. What happened in 1953? the big flood in the netherlands in retrospect. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 363(1831):1271–1291, 2005.

Tilmann Gneiting and Matthias Katzfuss. Probabilistic forecasting. *Annual Review of Statistics and Its Application*, 1:125–151, 2014.

Tilmann Gneiting, Adrian E. Raftery, Anton H. Westveld, and Tom Goldman. Calibrated Probabilistic Forecasting Using Ensemble Model Output Statistics and Minimum CRPS Estimation. *Monthly Weather Review*, 133(5):1098–1118, 2005. doi: 10.1175/MWR2904.1.

Floris Goerlandt and Jakub Montewka. Maritime transportation risk analysis: review and analysis in light of some foundational issues. *Reliability Engineering and System Safety*, 138:115–134, 2015.

Michael Hanss. *Applied fuzzy arithmetic*. Springer, 2005.

Gregory Hibbert, David O'Brien, et al. Berth to berth voyage schedule optimisation-a torres strait case study. In *Australasian Coasts and Ports 2019 Conference: Future directions from 40 [degrees] S and beyond, Hobart, 10-13 September 2019*, page 569. Engineers Australia, 2019.

Dominik Hose. The embarrassingly simple calculus of possibility theory. Risk Institute Online Talks, 2020. URL `https://riskinstitute.uk/riskinstituteonline/`.

Dominik Hose, Markus Mäck, and Michael Hanss. A possibilistic approach to the optimization of uncertain systems. *1*, 2018.

Dominik Hose, Markus Mäck, and Michael Hanss. Robust optimization in possibility theory. *ASCE-ASME J Risk and Uncert in Engrg Sys Part B Mech Engrg*, 5(4), 2019.

K David Jamison and Weldon A Lodwick. Minimizing unconstrained fuzzy functions. *Fuzzy Sets and Systems*, 103(3):457–464, 1999.

Elena Kelareva. The "DUKC Optimiser" ship scheduling system. In *2011 International Conference on Automated Planning and Scheduling System Demonstrations*, 2011.

Elena Kelareva, Sebastian Brand, Philip Kilby, Sylvie Thiebaux, and Mark Wallace. CP and MIP Methods for Ship Scheduling with Time-Varying Draft. In *Twenty-Second International Conference on Automated Planning and Scheduling*, ICAPS '12, 2012.

N Le Carrer, S Ferson, and P. L Green. Optimising cargo loading and ship scheduling subject to uncertain sea levels. 8th Workshop on Reliable Engineering Computing, 2018.

Noémie Le Carrer. Possibly extreme, probably not: Is possibility theory the route for risk-averse decision-making? *Atmospheric Science Letters*, page e01030, 2021.

Noémie Le Carrer and Scott Ferson. Beyond probabilities: A possibilistic framework to interpret ensemble predictions and fuse imperfect sources of information, 2020. Under review.

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

Noémie Le Carrer, Scott Ferson, and Peter L Green. Optimising cargo loading and ship scheduling in tidal areas. *European Journal of Operational Research*, 280(3):1082–1094, 2020.

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

Oleg Makarynskyy, D Makarynska, Michael Kuhn, and WE Featherstone. Predicting sea level variations with artificial neural networks at hillarys boat harbour, western australia. *Estuarine, Coastal and Shelf Science*, 61(2):351–360, 2004.

Marie-Hélène Masson and Thierry Denœux. Inferring a possibility distribution from empirical data. *Fuzzy Sets and Systems*, 157(3):319–340, 2006. ISSN 0165-0114. doi: https://doi.org/10.1016/j.fss.2005.07.007.

R Mel and Piero Lionello. Probabilistic dressing of a storm surge prediction in the adriatic sea. *Advances in Meteorology*, 2016, 2016.

Riccardo Mel and Piero Lionello. Storm surge ensemble prediction for the city of venice. *Weather and forecasting*, 29(4):1044–1057, 2014a.

Riccardo Mel and Piero Lionello. Verification of an ensemble prediction system for storm surge forecast in the adriatic sea. *Ocean Dynamics*, 64(12):1803–1814, 2014b.

K Mylne, C Woolcock, J Denholm-Price, and R Darvell. 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*, pages 113–118, 2002.

Terry O'Brien et al. Experience using dynamic underkeel clearance systems: selected case studies and recent developments. In *30th PIANC-AIPCN Congress 2002*, page 1793. Institution of Engineers, 2002.

Tim Palmer. The ecmwf ensemble prediction system: Looking back (more than) 25 years and projecting forward 25 years. *Quarterly Journal of the Royal Meteorological Society*, 145:12–24, 2019.

Pau Rabassa and Christian Beck. Superstatistical analysis of sea-level fluctuations. *Physica A: Statistical Mechanics and its Applications*, 417:18–28, 2015.

Adrian E. Raftery, Tilmann Gneiting, Fadoua Balabdaoui, and Michael Polakowski. Using Bayesian Model Averaging to Calibrate Forecast Ensembles. *Monthly Weather Review*, 133(5):1155–1174, 2005. doi: 10.1175/MWR2906.1.

David S Richardson. Skill and relative economic value of the ecmwf ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society*, 126(563):649–667, 2000.

Mark S. Roulston and Leonard A. Smith. Evaluating Probabilistic Forecasts Using Information Theory. *Monthly Weather Review*, 130(6):1653–1660, 2002. doi: 10.1175/1520-0493(2002)130<1653:EPFUIT>2.0.CO;2.

Mark S. Roulston and Leonard A. Smith. Combining dynamical and statistical ensembles. *Tellus A: Dynamic Meteorology and Oceanography*, 55(1):16–30, 2003. doi: 10.3402/tellusa.v55i1.12082.

Jin-Hwa Song and Kevin C Furman. A maritime inventory routing problem: Practical approach. *Computers and Operations Research*, 40(3):657–665, 2013.

George Sugihara, Robert May, Hao Ye, Chih-hao Hsieh, Ethan Deyle, Michael Fogarty, and Stephan Munch. Detecting Causality in Complex Ecosystems. *Science*, 338(6106): 496–500, 2012. ISSN 0036-8075. doi: 10.1126/science.1227079.

Burak Uslu, Andy Taylor, Greg Hibbert, Rafael Soutelino, et al. Connecting sea level forecasts with the bulk export industry. *Australasian Coasts and Ports 2017: Working With Nature*, page 1084, 2017.

L.A Zadeh. Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets and Systems*, 1 (1):3–28, 1978. ISSN 0165-0114. doi: https://doi.org/10.1016/0165-0114(78)90029-5.

Lotfi A Zadeh. The concept of a linguistic variable and its application to approximate reasoning—i. *Information sciences*, 8(3):199–249, 1975.

## Chapter 6

# Conclusion and Future works

## 6.1 Conclusions

We now summarize the findings of this PhD from the perspective of the research questions developed in the introductory Chapter 1, starting with:

**Q1** How can we optimise the cargo loading and ship scheduling decisions given imperfect sea level harmonic tide forecasts, without foregoing safety?

In Chapter 2, we have shown from realistic case studies between British ports that modelling the uncertainty on sea level residuals rather than using a rule-of-the-thumb safety margin was indeed promising in tidal areas. Fitting a classical distribution (e.g. logistic or mixture of Gaussian) to model the residuals in each port of call, and then combining a global optimisation algorithm (e.g. PSO) to Monte-Carlo sampling of these residuals altogether with a risk metric allows to improve the net benefit for the shipping company. Predictions are both robust and optimal, w.r.t. a standard safety margin. Besides, while the arbitrary safety margin can become close to optimality, achieving this requires (long) experience and remains problem dependent. On the contrary, such a stochastic modelling of the sea level residuals provides an optimal and robust solution from the first call. On top of that, we showed that even such a stationary modelling of sea-level residuals was robust to unseen extreme sea levels and remained in the latter case more optimal than a fixed safety margin.

Could we do better with actual, physical-model based sea level residual predictions? To that purpose, we started to investigate the concept of weather ensemble predictions. More precisely, we wondered whether, generally speaking, their probabilistic interpretation is the best way to extract the valuable predictive information they contain. Hence the following research questions, that we tried to answer in Chapters 3 and 4 by means of conceptual developments and associated numerical experiments on the Lorenz 96 system.

In these two works we investigate the following research questions:

**Q2a** 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?

**Q2b** Can we simultaneously maintain or improve the prediction skills compared to those of standard probabilistic interpretations?

**Q2c** How can we combine such a possibilistic framework with insights about the local dynamics of the system?

**Q2d** Can a possibilistic treatment of the EPS provide more formal guarantees than a probabilistic interpretation? If yes, at what cost?

**Q2e** Can we operationally use the possibilistic outputs at their full potential, that is more than simply deriving associated probabilities?

Chapter 3 develops a possibilistic interpretation of the EPS alone. It derives its properties in the case of a continuous reading of the resulting predictive possibility distribution, that is in terms of confidence intervals on the future value of the variable of interest. We compare the performances of these confidence intervals to the performances obtained by means of a classical probabilistic interpretation of EPSs. Therein, we show that our methodology provides confidence intervals associated with formal guarantees. The latter are verified experimentally also for extreme events, except for the asymptotic case of extremely large (EPS+observation) archives (which corresponds to the probability-like limit behaviour of possibility theory). In practice, such archives do not exist in the fast-changing field of ensemble forecasting models. We show that the confidence intervals based on our methodology overpass the probability-based ones in two cases: 1) at very small lead times for both common and extreme events, where they are as reliable yet narrower; 2) more blatantly, at intermediate and large lead times for extreme events, where they remain guaranteed and can be brought close to perfect reliability even for particularly rare events, yet at the expense of precision. These results can be reached with operational archive like the 20–30-year reforecast datasets. The guarantees are retained for smaller archives, which however lead to more conservative intervals and thereby impede operationality.

Chapter 4 continues the presentation of this approach by addressing two limitations of this first study, which 1) focused on the continuous reading of predictive distributions and consequently did not exploit at its full potential the possibilistic concepts of necessity, possibility and ignorance; and 2) did not take into account the local dynamics of the system (initial conditions), making EPS-based predictive distributions rather conservative.

In this chapter, we consequently investigated the benefits of aggregating two predictive possibility distributions, one from the EPS interpretation and one exploiting a past time series of the variable of interest including the initial conditions (i.e. time at which the prediction is made) by means of the analog method. We thus showed how the possibilistic framework allowed us to combine several incomplete sources of knowledge in a consistent manner, and thus to reduce their respective conservatism. Importantly, *our framework is more direct than the probabilistic one: we do not try to correct misleading EPS-based probabilities*, our outputs directly make sense without *a posteriori* calibration. Moreover, by using the credibility derived from our possibilistic framework as a probabilistic prediction, we are able to reproduce the classical 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 extreme events without deteriorating significantly the performances for other events (information content). Operationally, as could be expected, the performances of our possibilistic framework depend on the size of the archives at hand. In any case, *when it comes to extreme event 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 (the so-called "dynamical archive").

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 lies and allows us to go beyond standard probabilistic performances. In particular, we compared the effects of two aggregation methods, namely Zadeh's and the general aggregation method, and concluded that *Zadeh's was the most interesting trade-off between specificity and reliability of the resulting possibility distribution.* In the case of extreme events and large lead times, the risk-averse decision-maker may however prefer to use the general aggregation method if the dynamical archive at hand is not long enough to ensure the reliability of possibilistic outputs.

Different methodologies for emitting predictions from the possibilistic measures $N$ and $\Pi$ 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 both extreme and non extreme events simultaneously. We also discussed how to use at their full potential the couple $(N, \Pi)$: to derive predictive probabilities (e.g. by means of the credibility) and to estimate *a priori*, via the interval length $[N, \Pi]$ the trust we can have in their informativeness. The discussion on results and methodologies was again limited for shortness of space and time, however it opens the way for further investigation about how to best use these measures in an operational context.

If we 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.* Besides, 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 when interpreted as a probability. Possibility theory seems consequently an interesting alternative to the classical probabilistic interpretations of EPSs, given the same sources of information (EPS archives). Other imprecise probability frameworks such as credal sets could have been investigated, however we believe *possibility theory is more interesting for its intuitive rationale (especially in a phenomenological context like the weather system where explanatory power, "making sense" is important) and potential for communication to the end-users, as well as for its simplicity and power when it comes to fusing information from various sources.*

To summarize, these results show that such a possibilistic framework allows to extract globally as much information from the EPSs as a classical probabilistic treatment would do and sometimes more (large lead times and extreme event prediction). This formulation can seem counter-intuitive, given that the possibilistic representation is known as *weaker* [Dubois et al., 1993] than the probabilistic one. However this weakness comes from the fact that imprecision is taken into account in the possibilistic modelling. In situations of incomplete information, typically like the EPSs given the limitations raised in Section 1.1.2 about the way they are produced, possibility theory allows to extract more information than probability theory.

We consequently tried to apply it in a real-world situation, namely the shipping optimisation problem introduced in Chapter 2. Thus the final Chapter 5 studies the research questions:

**Q3a** How valuable is the information extracted from the storm-surge EPS, either via a probabilistic approach or via a possibilistic approach, for an application such as maritime shipping optimisation?

**Q3b** In particular, is this information more valuable for this specific application than a classical Monte-Carlo-based optimisation using harmonic tide predictions and historical best-fit modelling of sea level residuals in each port?

The results presented in that chapter confirm, when it comes to the predictive performance of the possibilistic interpretation of EPS, that the size of the archive

(EPS-observation) at hand matters for such a similarity-based approach. For very small archives, a probabilistic interpretation will generally work equally well or better, at the exception of extreme events. This finding confirms what we found in the above-mentioned previous works, although it is not as marked in this study due to the very small test dataset at hand ($52$ versus $40.10^3$ in Le Carrer [2021]).

Regarding the application however, we found that, *when it comes to using EPS as source of information on sea levels, only a possibilistic interpretation led to reasonable decisions w.r.t. ship routing*. The probabilistic approach, due to its lack of conservatism (contrary to the frequent classification of predictions in the 'ignorance' area in the possibilistic case), tends to lead to biased predictions and decisions that are consequently almost sure fails, as our initial experiments revealed in the introduction of Chapter 3.

Beyond that and to answer our initial question, this small experiment shows that the most optimal way of taking scheduling decisions among our propositions is a simple optimisation based on static, historical best-fit modelling of sea level residuals. This equals to adding a rule-of-the-thumb safety margin to the ship's draft, without the need to find out this problem-dependent safety margin (also less robust to the non-stationarity of sea level distributions). Our possibilistic approach, in these particular conditions of extremely limited archive, is too conservative (ignorance prevails) to compete with the probabilistic modelling above-mentioned. However, the analysis of the effect of parameter $\beta$, namely the confidence level in constraint satisfaction, suggests that a careful choice (here $\beta = 0.95$) already allows to achieve decisions sometimes challenging those given by the probabilistic method. This is promising for similar works with larger EPS archives.

## 6.2 Overall contribution and future works

At a very high level, the novelty and major contributions of this PhD are :

(i) A risk analysis of cargo loading and ship scheduling decisions in tidal areas and two methodologies to provide robust and optimal decisions given either purely deterministic tide predictions (with an archive of the observed residuals) or numerical surge ensemble predictions in each port of call (with EPS archives in the same ports of call and possibly long past time-series of the surges) ;

(ii) A data-driven possibilistic framework to interpret weather ensemble predictions (the first to the knowledge of the authors) instead of the classical probabilistic approach and its analysis and justification, in particular for the prediction of extreme events ;

(iii) An application of possibility theory in practical fields ( (shipping) global optimisation and weather forecasting), which is interesting for highlighting the potential advantages and drawbacks of such an alternative approach w.r.t. more classical, mainstream probability-based methodologies.

The practical implications of our work are:

(a) Raising awareness of the economic potential of taking into account sea level uncertainty in scheduling decisions more finely than a 'rule-of-the-thumb' safety margin, not only for the more studied expensive freight and large ships but also for the masses of small vessels (mini-bulkers), cheap commodities (grains) and small ports strongly affected by tidal effects (i.e. limited dredging), which in the current context of transportation greening may be a non-negligible lever of progress;

(b) Contributing to the diffusion and application of possibility theory to applied fields, here weather forecasting, to show the interest of going beyond "classical probabilities" when the latter do not always make sense or are not the most adapted, e.g. for extreme events predictions in the ensemble forecasting context.

These two investigations could be completed in many ways, including:

**When it comes to the shipping optimisation problem:**

1. One way to go would have been to study in depth the probabilistic modelling and refine it by e.g. taking into account the space-time dependence between residuals from different locations and/or time, in particular its cyclic nature (due to tides), or modeling tide residuals differently (e.g. neural networks [Pashova and Popova, 2011], superstatistics [Rabassa and Beck, 2015]).

2. Another way would have been to work on the efficient optimisation side of the problem and, instead of a simple double-loop Particle-Swarm-Optimisation where a Monte-Carlo sampling is nested, provide a faster optimisation algorithm that scales well with the number of ports at hand.

**When it comes to possibility theory for interpreting EPSs:**

1. It would be interesting to see whether the conclusions obtained on the Lorenz 96 toy system apply to real-world weather EPS.

2. As raised by one of the reviewers of the study presented in Chapter 3, in practice the verification (as observation) is a random variable itself . The use of confidence intervals rather than a Bayesian formalism and the derivation of credible intervals may consequently be discussed. Since our approach is taking such impreciseness into account (limited volume $S_{x_t}$ around $x_t$, Masson and

Denoeux's transformation – cf. Section 3.3.1), even without explicitly tackling this problem, our framework accounts for (reasonable) randomness in the so-called verification. However, it could be interesting to take this explicitely into account.

3. We assumed that the dynamical system at hand was close to stationarity. It would be worth running similar experiments with other non-stationary toy systems, to assess if possibilistic framework is more robust than a purely probabilistic one.

4. The analysis of how to use best the dual possibility and necessity measures depending on prediction strategies (risk-averse, risk-prone, etc) that we presented here was a preliminary result. There is much room for extension and presenting proper examples of applications.

5. Finally, to assess the skills of the our possibilistic predictions, we often issued derived probabilities and used the standard metrics to assess the predictive performances of probabilistic predictions. It would be worth investigating how to develop scores fitted to possibilistic outputs directly, which would also allow to make more sense of these outputs (without turning back to probabilities) and improve the spread and communication around possibility theory in applied fields.

**When it comes to the use of surge EPSs in shipping optimisation:**

1. Given the lack of real data characterizing this work, repeating the experiment with a larger archive of EPS predictions (typically 2-3 years instead of 6 months, cf. experiments presented in Le Carrer [2021]) would consequently be an interesting future work. We expect an improvement of the average possibility-based results, which means that those with well-chosen level of constraint satisfaction $\beta$ could challenge more seriously the probabilistic outputs.

2. Beyond that, another interesting work would be to compare the effect of the number of ports in the decision-making problem on these comparative performances. Taking into account several ports at a time increases the conservatism of the possibility distribution accounting for the satisfaction of all constraints [Hose, 2020]. This could be a serious limitation with respect to the probabilistic approach who suffers less of this asymptotic behaviour.

# Bibliography

Jochen Bröcker and Leonard A. Smith. From ensemble forecasts to predictive distribution functions. *Tellus A: Dynamic Meteorology and Oceanography*, 60(4):663–678, 2008. doi: 10.1111/j.1600-0870.2007.00333.x.

Didier Dubois, Henri Prade, and Sandra Sandri. On Possibility/Probability Transformations. In R. Lowen and M. Roubens, editors, *Fuzzy Logic: State of the Art*, pages 103–112. Springer Netherlands, Dordrecht, 1993. ISBN 978-94-011-2014-2. doi: 10.1007/978-94-011-2014-2_10.

Dominik Hose. The embarrassingly simple calculus of possibility theory. Risk Institute Online Talks, 2020. URL https://riskinstitute.uk/riskinstituteonline/.

Noémie Le Carrer. Possibly extreme, probably not: Is possibility theory the route for risk-averse decision-making? *Atmospheric Science Letters*, page e01030, 2021.

Lyubka Pashova and Silviya Popova. Daily sea level forecast at tide gauge burgas, bulgaria using artificial neural networks. *Journal of sea research*, 66(2):154–161, 2011.

Pau Rabassa and Christian Beck. Superstatistical analysis of sea-level fluctuations. *Physica A: Statistical Mechanics and its Applications*, 417:18–28, 2015.

## Appendix A

# Optimising cargo loading and ship scheduling subject to uncertain sea levels - Risk models

We report here an extract of the conference paper Le Carrer et al. [2018], that was later extended into the journal article presented in Chapter 2. In this conference paper, we developed more the question of risk models (i.e. how to model risk through an objective function to be optimised during the optimisation procedure) and presented their performances and robustness in a case study similar to the $N = 2$ - ports case study presented in Chapter 2, Section 2.2.1. The context, type of ship and commodity are the same, as well as mathematical notations. Only ports differ. On November 19th 2016 at 16:30 UTC, we assume that the shipper has to decide how much barley will be freighted and when the vessel will depart from Lowestoft Port to Portsmouth Harbour, both on the British coast. To this purpose, they use the long term harmonic tide forecasts as sea level predictions as well as the decision model described in Sections 2.2.2 and 2.2.3.

## A.1 A probabilistic approach to decision making

Using the model described above, one can choose an optimisation technique (e.g. particle swarm optimisation or simulated annealing) to compute the optimal decision to take at time $t_0$, according to the sea level forecast time series $\hat{X}_p(t)$ for the two ports $p = \{p_1, p_2\}$. Such a calculation does not consider the actual stochastic behaviour of the water depth. Mean sea levels are locally influenced by a range of factors, including weather. A residual $e_p(t) = X_p(t) - \hat{X}_p(t)$ between the predictions and the observations can lead to either a regret ($e_p > 0$: the shipper could have loaded more or departed earlier) or a loss ($e_p < 0$: in order to adjust to the actual water level the journey is delayed, or a grounding can happen). In other words, the resulting solution is risky as it does not tolerate a negative deviation to prediction nor port delays. In order to account for the uncertainty on the outcome of a given decision and its potentially dramatic consequences for the shipping company, it is sensible to work in the frame of risk averse optimisation.

From the classical mean-risk [Markowitz, 1952] and chance-constrained perspectives [Charnes et al., 1958] to the more recent so-called robust optimisation models (e.g. worst-case, minimax regret, uncertainty sets, see Greenberg and Morrison [2008] for an historical overview and Shapiro et al. [2009] for an extensive presentation), operational research has developed a range of approaches to address the notion of uncertain decision-making. In these problems, the questions at stake are: a) Are all the scenarios acceptable, or feasible, whatever their probability of occurence? (e.g. is a ship grounding acceptable?) b) How much does the decision-maker give way to objective optimality in order to guarantee feasibility? Any solution to stochastic optimisation is a trade-off between feasibility and performance, or said otherwise, between variance and guaranteed value of the objective function.

A (robust) optimisation approach must thus define the attitude of the decision-maker towards risk and the specificities of her optimisation problem before computing any solution. Between the two extreme approaches that are worst-case (always feasible) and deterministic optimisation (best performance e.g. for the most probable scenario, no uncertainty taken into account), lie a range of models depending on the decision-maker's requests as regards performance and feasibility. We introduce in the following a representative selection of them, before comparing their outputs in Section A.2.

### A.1.1 Risk models

#### A.1.1.1 Regret

In decision-making under uncertainty, it is common to adopt the gain shortfall perspective. In this case, risk takes the meaning of the loss in profit due to the fact that decision $\boldsymbol{d} \in \mathcal{D}$ is taken at time $t_0$ based on imperfect forecasts $\hat{X}_p \in \mathcal{X}$ of the environment state $X_p \in \mathcal{X}$. Let $F_p$ be the cumulative distribution function over $X_p$, which is conditional on information on the prior values of $X_p$ and possible other information. Let $\hat{F}_p$ be a predictive distribution of $X_p$ (that is a distribution over $\hat{X}_p$) provided by the forecaster at $t_0$. Let $\hat{X}_p(t)$ be a point forecast time series of $X_p(t)$ over time $[t_0, t_0 + T]$, $B(.,.) : \mathcal{D} \times \mathcal{X} \rightarrow \Re$ the utility function (namely the net benefit of the journey based on decision $\boldsymbol{d}$) and $y(\cdot) : \mathcal{X} \rightarrow \mathcal{D}$ an optimal action function defined by:

$$y(\hat{F}_p) = \underset{\boldsymbol{d} \in \mathcal{D}}{\arg\max} \left( \mathbb{E}[B(\boldsymbol{d}, \hat{X}_p)]_{\hat{F}_p} \right) = \underset{\boldsymbol{d} \in \mathcal{D}}{\arg\max} \int_{\mathcal{X}} B(\boldsymbol{d}, \hat{X}_p) d\hat{F}_p \qquad \text{(Equation 1)}$$

The loss function $L(.,.) : \mathcal{D} \times [0, 1] \rightarrow \Re$ is then defined by Granger and Machina [2006] as:

$$L\Big(y(\hat{F}_p), F_p\Big) = B\Big(y(X_p), X_p\Big) - B\Big(y(\hat{F}_p), X_p\Big) \qquad \text{(Equation 2)}$$

for all $\hat{X}_p, X_p \in \mathcal{X}$. In other words, the utility of the decision made under uncertainty $B\Big(y(\hat{F_p}), X_p\Big)$ is compared to the utility resulting from the decision made under perfect knowledge of the future $B\Big(y(X_p), X_p\Big)$.

With an absolute robust approach, each possible shipping decision $\boldsymbol{d}$ is mapped to the maximum loss it can generate, whatever its probability of occurrence. The optimal decision minimises:

$$\boldsymbol{d}^{*} = \min_{\boldsymbol{d}\in\mathcal{D}} \left\{ \max_{F_p} \left\{ L\Big(\boldsymbol{d}, X_p\Big) \right\} \right\} \tag{Equation 3}$$

Its less conservative counterpart involves mapping each decision $\boldsymbol{d}$ to the regret it generates in average:

$$\boldsymbol{d}^{*} = \min_{\boldsymbol{d}\in\mathcal{D}} \left\{ \mathbb{E}\left[ L\Big(\boldsymbol{d}, X_p\Big) \right]_{F_p} \right\} \tag{Equation 4}$$

Looking more closely at the definition of the loss which we aim to minimise (the expectation over the space of sea level residuals), one can notice that minimising $\mathbb{E}\left[ L\Big(\boldsymbol{d}, X_p\Big) \right]_{F_p}$ is equivalent to finding the decision $\boldsymbol{d}^{*}$ that maximises the expected benefit $\mathbb{E}\left[ B\Big(\boldsymbol{d}, X_p\Big) \right]_{F_p}$.

A.1.1.2 Mean-risk

What appears to be the first risk model developed in operational research involves adding a penalty known as the risk functional to the expected objective outcome of a given decision, and thus setting:

$$\boldsymbol{d}^{*} = \max_{\boldsymbol{d}\in\mathcal{D}} \left\{ \mathbb{E}\left[ B\Big(\boldsymbol{d}, X_p\Big) \right]_{F_p} - \beta \mathbb{D}[B]_{F_p} \right\} \tag{Equation 5}$$

where the parameter $\beta \geq 0$ allows to quantify the price of risk.

In the simplest case, the risk functional is proportional to the standard deviation of the objective:

$$\mathbb{D}[B] = \Big( \mathbb{E}\left[ (B - \mathbb{E}[B])^2 \right]_{F_p} \Big)^{1/2} \tag{Equation 6}$$

Negative and positive deviations to the mean do not have the same implications in terms of risk. In the case of maximising the shipping benefit, positive deviations to the expected benefit are welcome, contrary to negative ones. The standard deviation cannot fully describe such assymetrical behaviour of the utility function. The lower semi-deviation of order $\gamma$ is consequently introduced as:

$$\mathbb{D}[B] = \Big( \mathbb{E}\left[ (B - \mathbb{E}[B])_{-}^{\gamma} \right]_{F_p} \Big)^{1/\gamma} \tag{Equation 7}$$

Note that, in the following, we use $\gamma = 2$ and $\beta = 1$.

#### A.1.1.3 Worst-case

The absolute robust way of optimising the shipping net benefit is to prevent any unfeasible scenario and maximise the outcome in the worst possible scenario. In other words, finding the decision:

$$\boldsymbol{d}^* = \max_{\boldsymbol{d} \in \mathcal{D}} \left\{ \min_{F_p} \left\{ B\Big(\boldsymbol{d}, X_p\Big) \right\} \right\} \qquad \text{(Equation 8)}$$

#### A.1.1.4 Chance-constrained

Although strictly speaking robust in terms of feasibility, the worst-case approach is often criticised for being too conservative in practical implementations.

The chance-constrained perspective allows more flexibility. Given a level of guarantee $\zeta$, it computes the decision maximising the ensured benefit at this level, in other words:

$$\boldsymbol{d}^* = \max_{\boldsymbol{d} \in \mathcal{D}} \left\{ \inf_{b} \left\{ P\left( B\Big(\boldsymbol{d}, X_p\Big) \leq b \right) \leq 1 - \zeta \right\} \right\} \qquad \text{(Equation 9)}$$

In our experiments, we use $\zeta = 0.98$, that is we look for the maximal benefit allowing an error rate less than or equal to 1%.

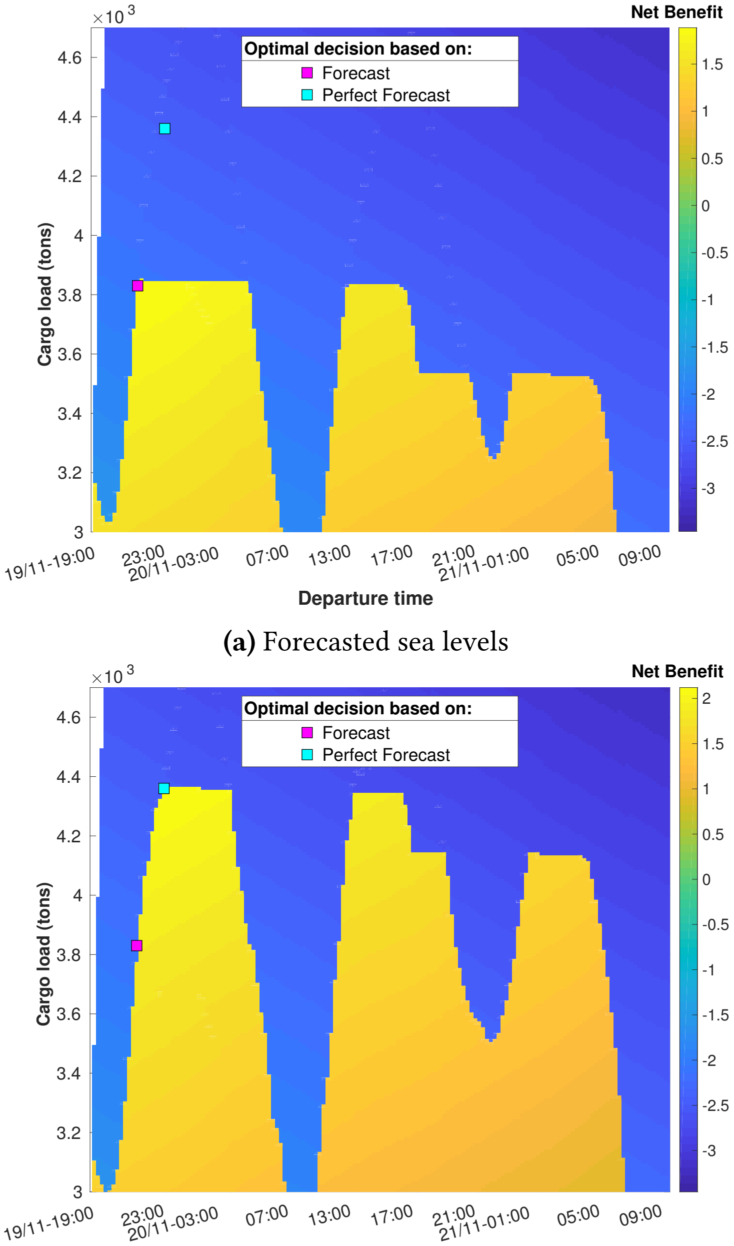
## A.2 Results and Discussion

All the results in terms of benefit $B$ will be expressed as multiples of the value of the minimum cargo load, $B_0 = \text{US\$ } 363,550$. We also set the cost of not making the delivery in time to ($Z = -V - (O + P + U)$). Negative benefits would thus imply a grounding or the impossibility to reach the arrival port within the specified time horizon.

### A.2.1 Deterministic case

The $B_{PSO}$ procedure recommends the ship to leave Lowestoft Harbour at 23:00 UTC on November 19th 2016 with an overall barley freight of $3,835.0$ mt. The standard deviations of these recommendations are estimated to be $0.5$ mt in freight and less than $15$ minutes in time (from 1,000 independent runs).

Figure A.1 presents a mapping of the final shipping benefit over the decision search space $\mathcal{D}$, given the forecast *a priori* at hand and given perfect forecasts, i.e. the *a posteriori* exact observations of the sea level depths. The optimal decision according to $B_{PSO}$ in each scenario differ by 1 hour and 30 minutes in time and $527$ mt in cargo load. In other words, the deterministic solution under imperfect harmonic predictions is far away from optimality in the real-world of non-zero residuals. Besides, it is



**(a)** Forecasted sea levels

**(b)** Actual sea levels

**Figure A.1** Mapping of the net benefit $B$ over all the decisions $(t_d, m)$ of the search space, given sea level forecasts at hand (a) or actual sea level (b). The optimal decisions based on the deterministic forecasts and on the perfect forecasts (i.e. real state of the sea) through the solver $B_{PSO}$ are also reported.

**Table A.1** Statistics over 50 runs of the outputs in terms of decision-making. The optimal cargo load $m$, departure time $t_d$ and guaranteed benefit $B_{.98}$ at the level of 2 % (over 100,000 simulations) are expressed in metric tons, UTC and fraction of $B_0$ respectively. The uncertainty is computed as the standard deviation of the results.

| **Distribution**<br>**Risk metric** | **GMM** | **Logistic** | **Gaussian** |
|---|---|---|---|
| Mean-Regret | $m = 3,947 \pm 20$<br>$t_d = 00:45 \pm 30$mn<br>$B_{.98} = 1.901$ | $m = 3,940 \pm 10$<br>$t_d = 01:00 \pm 15$mn<br>$B_{.98} = 1.894$ | $m = 3,957 \pm 9$<br>$t_d = 00:30 \pm 15$mn<br>$B_{.98} = 1.909$ |
| Worst-Case | $m = 3,943 \pm 15$<br>$t_d = 00:45 \pm 15$mn<br>$B_{.98} = 1.899$ | $m = 3,935 \pm 10$<br>$t_d = 01:00 \pm 30$mn<br>$B_{.98} = 1.891$ | $m = 3,961 \pm 11$<br>$t_d = 00:45 \pm 30$mn<br>$B_{.98} = 1.908$ |
| Mean-Risk | $m = 3,946 \pm 18$<br>$t_d = 00:45 \pm 30$mn<br>$B_{.98} = 1.901$ | $m = 3,933 \pm 15$<br>$t_d = 00:45 \pm 15$mn<br>$B_{.98} = 1.895$ | $m = 3,963 \pm 16$<br>$t_d = 01:00 \pm 30$mn<br>$B_{.98} = 1.905$ |
| Chance-Constrained | $m = 3,959 \pm 11$<br>$t_d = 00:30 \pm 15$mn<br>$B_{.98} = -2.239$ | $m = 3,956 \pm 9$<br>$t_d = 00:45 \pm 15$mn<br>$B_{.98} = 1.905$ | $m = 3,976 \pm 6$<br>$t_d = 00:45 \pm 15$mn<br>$B_{.98} = -2.253$ |

quite straightforward to see on these maps that both solutions are very sensitive to perturbations. A 15 mn departure/arrival shift or a negative error in the actual sea levels both shift the expected benefit from its maximum to the negative area.

One way to get over the second limitation is to improve the accuracy of sea level forecasts. This is currently achieved by means of storm surge models. To take into account the local weather perturbations, these models use atmospheric forecasts as forcing in shallow-water hydrodynamic simulations e.g. the CS3 storm surge model covering the sea of the northwest European continental shelf [Flowerdew et al., 2010]. Nevertheless, whatever the accuracy reached, these forecasts cannot prevent the issue of port perturbations and delays. Hence it seems reasonable to develop a robust solution instead of a single deterministic optimisation.

### A.2.2 Risk models

We now use $R_{PSO}$ to compute the optimal shipping decision under uncertain sea levels. Each of the four risk metrics presented in Section A.1.1 is combined with one of the three sea level residuals distribution models under consideration. Table A.1 reports the statistical results of each combination as regards the optimal cargo load, departure time and the resulting guaranteed benefit at the error rate of 2 %, that is the 2% percentile $B_{.98}$. The latter is estimated from 100,000 Monte Carlo simulations. In order to prevent a methodological bias, these simulations sample the sea level by means of bootstrapping (over dataset $D_u$, c.f. Section 2.4.0.1).

As the purpose of the $R_{PSO}$ procedure is to support decision-making, it is necessary to analyse the consequences of the above results as regards their translation in terms of practical shipping decision. The overall majority of the computed departure times are located within a 30 mn time slot centered on 00:45. Taking into account the relative inertia of large vessels and generally slow port dynamics (from decision to subsequent actions), this range of uncertainty can be seen as a buffer to consider in the decision-making schedule. Trying to increase the precision on $t_d$ would be meaningless considering the real world context of a maritime shipping problem.

As expected, the worst-case approach is the more conservative and generally computes the lowest loads ($m \approx 3,945$ mt overall). The mean-risk model with a penalty equal to one standard lower deviation behaves very similarly. The chance-constrained approach returns the highest loads ($m \approx 3,966$ mt overall) and the mean-regret (or expected-benefit) approach is intermediary. This is a quite general observation, whatever the sampling distribution. As regards the distribution impact, Logistic sampling produces more conservative loads than the GMM approach and further again, than the Gaussian one. The difference between the maximal and minimal loads abovementioned is in the range of $30$ mt, that is in our case study less than 3 centimetres of draft. This invariably leads to quite similar guaranteed benefits $B_{.98}$ between the worst-case, mean-risk and mean-regret for a given distribution. On the contrary, the guaranteed benefit of the chance-constrained solution is much less stable: either maximum or minimum (and) negative. This illustrates the concept of distributional (non-) robustness: according to the sea level modelling (Logistic versus Gaussian and GMM), the solution computed by $R_{PSO}$ leads to either very satisfying outcomes overall or to a very likely failure.

Figure A.2 summarises most of the information discussed above: whatever the risk metric, a Logistic sampling will produce more stable (smaller variance) outcomes than the other models. It also shows that, strictly speaking, only the mean-risk approach could be said to be distributionally robust. Indeed, the ranges of the reduction in standard deviation and in guaranteed benefit when the underlying distribution varies (2 and 0.6 % respectively) are much smaller than for the other metrics (closer to 8 and 0.9 % respectively). Considering the money at stake, even variations of 0.1% $B_{.98}$ are worth a few thousand dollars, so should not be neglected. Three observations can be highlighted as well. First, in this particular case study, the stochastic optimisation based on risk metrics allows the owner to (in most of the configurations) save money as the guaranteed benefit is above the expected benefit of the deterministic decision in real conditions. Second, the spatial organisation of the points underlines a general pattern in robust optimisation: the guaranteed benefit increases at the cost of the increase in variance [Gotoh et al., 2015]. Finally, as noted by Gotoh et al. [2015], the variation in actual benefit is about one order of magnitude smaller than the reduction in its standard deviation.

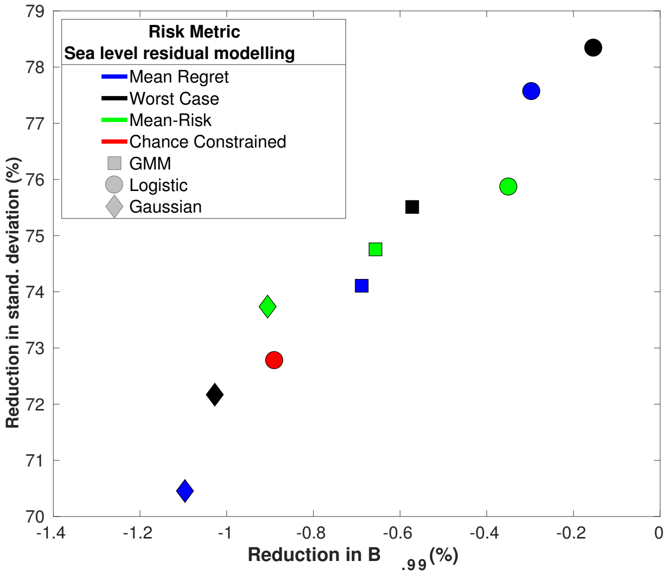


**Figure A.2** Performance of each optimisation approach (a risk metric combined with a sea level residuals distribution) from the perspective of the reduction of the guaranteed benefit at the error level of 2% and the standard deviation of the actual shipping benefit, with respect to the performances of the "deterministic" solution based on sea level forecasts alone. 100,000 Monte Carlo simulations are used to compute these statistics, with bootstrap sampling. The chance constrained and GMM or Gaussian sampling are not represented here as the reduction in guaranteed benefit is out of scope, reaching 200%.

## A.3 Conclusion

Figure A.3 summarises some of the above considerations in a 3-dimensional view of the optimisation problem. A map of the standard deviation is estimated with bootstrap sampling for each couple $(t_d, m)$ of the search space. On top of the map, we report the decision suggested by the net benefit optimisation from sea level forecasts, perfect forecasts (i.e. perfect knowledge of the future) and by four optimisation approaches. Figure A.3 gives a good overview of the set of solutions returned by all the approaches and presented above.

As the owner of the company, you could use the benefit optimisation decision that is based on the deterministic harmonic forecasts, load $3,835$ mt of barley and cast off at 23:00. However the outcome of this decision, given the actual observations of sea levels is $-2.15B_0$. This is much less desirable than the benefit $2.12B_0$ that you could make if you knew the future perfectly and left Lowestoft port at 00:30 with $4,362$ mt on board. Using the stochastic optimisation method developed in this paper, you could load cargo between $3,935$ and $3,959$ mt, raise anchor between 00:30 and 01:00 and get a net benefit from $1.89B_0$ to $1.91B_0$. If these decisions were reported in Figure A.1(b) (mapping based on actual sea level conditions), one could notice that a port re-scheduling of up to 2 hours (earlier) or 4 hours (delay) would

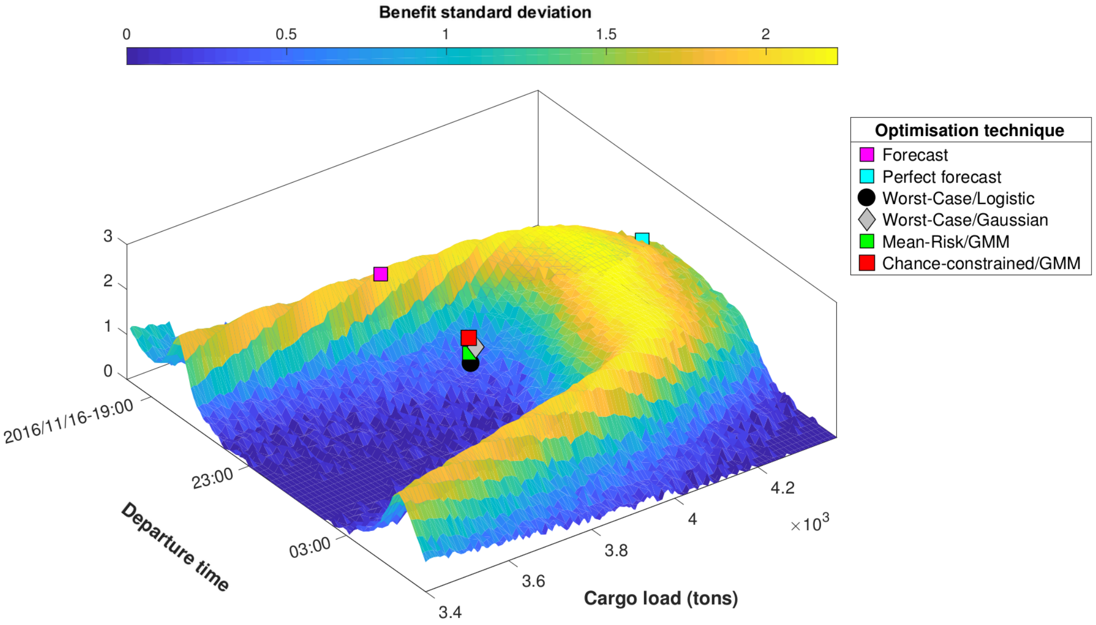


**Figure A.3** Three dimensional mapping of each decision $(t_d, m)$ to the associated actual benefit standard deviation. Points of interest discussed in the text are also reported. The mapping use Monte Carlo simulations of 1,000 journeys by means of bootstrap re-sampling.

not substantially change the benefit, nor a variation (in standard limits) in sea level conditions. Besides, Figure A.3 reminds that the variance in the actual benefit is substancially reduced for our solutions, contrary to the variance of the deterministic proposition. In other words, the approach $R_{PSO}$ proposes a robust solution. This is true for any risk metric introduced here apart from the chance-constrained, and true for any sampling distribution although a Gaussian generally leads to solutions with less predictible economic outcomes. Recalling the questions raised in the motivation of the problem (Section 2.1), in this case study, our stochastic approach demonstrated to be economically valuable with respect to the standard (deterministic) approach. Besides, a simple Logistic modelling of the residuals is enough to produce quality results, similar to those gained by means of a GMM.

One can note that the cargo load output $m^*$ can be turned into a safety margin $\Delta r$ to be deducted from the maximum draft that would have been allowed given the sea level tide forecasts at hand at $t_0$ (procedure $B_{PSO}$). For future works, it would be interesting to compare $\Delta r$ with what "non-stochastic" commercial softwares would suggest on a similar problem, so as to assess the quality and potential added value of our model.

Avenues of research on the problem raised in this paper include defining sounder uncertainty sets on which the risk metrics would then be applied. A finer modelling of the sea level residuals would also be judicious, exploiting the cyclic character of data.

# Bibliography

Abraham Charnes, William W Cooper, and Gifford H Symonds. Cost horizons and certainty equivalents: an approach to stochastic programming of heating oil. *Management Science*, 4(3):235–263, 1958.

Jonathan Flowerdew, Kevin Horsburgh, Chris Wilson, and Ken Mylne. Development and evaluation of an ensemble forecasting system for coastal storm surges. *Quarterly Journal of the Royal Meteorological Society*, 136(651):1444–1456, 2010.

Jun-ya Gotoh, Michael Jong Kim, and Andrew Lim. Robust empirical optimization is almost the same as mean-variance optimization. 2015. Accessed:2018-05-10.

Clive WJ Granger and Mark J Machina. Forecasting and decision theory. *Handbook of economic forecasting*, 1:81–98, 2006.

H.J. Greenberg and T. Morrison. *Robust optimization, Operations Research and Management Science Handbook.* CRC Press, Boca Raton, Florida, 2008.

N Le Carrer, S Ferson, and P. L Green. Optimising cargo loading and ship scheduling subject to uncertain sea levels. 8th Workshop on Reliable Engineering Computing, 2018.

Harry Markowitz. Portfolio selection. *The journal of finance*, 7(1):77–91, 1952.

Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczyński. *Lectures on stochastic programming: modeling and theory.* SIAM, 2009.

## Appendix B

# A (parametric) possibilistic framework for EPS interpretation

In this Appendix, we report our first tentative of possibilistic framework to interpret weather ensemble forecasts. It was published in the proceedings of the 2019 Annual Meeting of the European Meteorological Society [Le Carrer and Green, 2020]. For conciseness, we only report the framework and the empirical evaluation of its performances. The contributions of the authors are the following: NLC conceived of the presented idea, designed and implemented the research and wrote the article. PG contributed to the analysis of the results and reviewed the article.

## B.1 Possibilistic framework for EPS interpretation

The statistical post-processing of EPS generates forecasts in the form of predictive probability distributions $p(x|\tilde{\boldsymbol{x}}, \theta)$, noted $p(x|\tilde{\boldsymbol{x}})_\theta$, where $\tilde{\boldsymbol{x}} = \{\tilde{x}_1, \ldots, \tilde{x}_M\}$ is the ensemble, $\theta$ a vector of parameters and $p$ a (sum of) parametric distribution(s). BMA distributions 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 comes down to an ensemble dressing procedure. We compare our method against a Gaussian ensemble dressing, whose predictive probability distribution reads:

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

where $\mathcal{N}(\mu, v)$ is the normal distribution of mean $\mu$ and variance $v$. The parameters $\theta = \{a, \omega, \sigma\}$ are inferred through the optimization of a performance metric, e.g. the ignorance score [Roulston and Smith, 2002], or negative log-likelihood, a strictly proper[1] and local[2] logarithmic score.

[1] i.e. it takes its optimal value only when the forecast probability is equal to the true distribution of the system.

[2] i.e. it does not depend on the full forecast distribution, but only on the predictive probability associated to the true system's state.

Here, instead of performing a probabilistic ensemble dressing, we can perform a *possibilistic* ensemble dressing: a possibilistic membership function is dressed around each ensemble member first shifted and scaled. Similarly to its probabilistic twin, the $i$th possibility kernel is assumed to represent the possibility distribution of the true state of the system, given the observation of $\tilde{x}_i$. Because we have several member observations $i = \{1, \ldots, M\}$ and there is only one truth (the actual system's state), we can interpret it as a union (OR) of possibilities. Fuzzy set theory offers several definitions for computing the distribution resulting of the union of two fuzzy distributions. We adopt here the max-sum definition: $\pi_{A \cup B}(x) = \max\big(\pi_A(x), \pi_B(x)\big)$, although some of our tests, not presented here, show that alternative definitions do not significantly change results.

Gaussian kernels $\exp^{-\frac{1}{2}u_i^2}$ are thus fitted to each member $\tilde{x}_i$, with $u_i = \frac{x-(a\tilde{x}_i+\omega)}{\sigma}$, $a$ the scaling factor, $\omega$ the shifting of the kernels' peaks from the individual member $\tilde{x}_i$ and $\sigma$ a parameter accounting for the width of the individual kernels. The resulting possibilistic distribution is given by the sum, in a possibilistic manner, of all the individual kernels:

$$\pi(x) = \bigcup_{i=1\ldots M} \exp^{-\frac{(x-(a\tilde{x}_i+\omega))^2}{2\sigma^2}} = \sup_{i=1\ldots M} \exp^{-\frac{(x-(a\tilde{x}_i+\omega))^2}{2\sigma^2}} \qquad \text{(Equation 2)}$$

For any event of interest $A = \{x \in S_A\}$, we can extract the possibility and necessity measures $\Pi(A, \theta)$ and $N(A, \theta)$ (noted $\Pi_\theta(A)$ and $N_\theta(A)$), given the knowledge encoded in $\pi(x, \theta)$ (noted $\pi_\theta$). $\Pi_\theta(A)$ evaluates to what extent $A$ is logically consistent with $\pi_\theta$ whereas $N_\theta(A)$ evaluates to what extent $A$ is certainly implied by $\pi_\theta$. Ideally, this pair falls in an area of the possibilistic diagram $(N, \Pi)$ that is close to one of the three notable points: $(1, 1)$ for $A$ certain; $(0, 0)$ for $\bar{A}$ certain; $(0, 1)$ for total ignorance, i.e. both $A$ and $\bar{A}$ are possible but none is necessary given $\pi$. Points on the line $N = 0$ are in favor of $\bar{A}$, the more favorable the closer to $(0, 0)$; points on the line $\Pi = 1$ are in favor of $A$, the more favorable the closer to $(1, 1)$. Other areas of the diagram are inconsistent with the axioms defining $\Pi$ and $N$.

From the geometric interpretation given by the possibilistic diagram, several options are available for scoring each point $\big(N_\theta(A), \Pi_\theta(A)\big)$ that is, for assessing the quality of the prediction given by the pair $\big(N_\theta(A), \Pi_\theta(A)\big)$. A brute-force method is to minimize the distance to the correct pole (e.g. $(1, 1)$ for $A$ true). Yet, such an approach would try and push events towards $(1, 1)$ or $(0, 0)$ on the possibilistic diagram, thus ignoring the ignorance pole and, as a result, the idea that some events are impossible to predict from a particular EPS set. A more complete method could, for instance, also consider the rank $r$ of the EPS w.r.t. $A$. Namely, if the actual observation $x^*$ is in $S_A$, the associated point should belong to the line $\Pi = 1$ but the distance to the ignorance pole $(1, 0)$ should be proportional to $r$. The same applies for $x^* \notin S_A$; the associated point should belong to line $N = 0$ with the distance to $(1, 0)$ proportional to $r_{\bar{A}} = M - r_A$. Thus, an observation $x^* \in S_A$ associated to an erroneous EPS ($r \to 0$) will fall close to the ignorance pole, suggesting that we cannot trust the raw ensemble. A score verifying these requirements is:

$$S_i(\theta) = \begin{cases} |N_\theta(A) - \frac{r}{M}| + |\Pi_\theta(A) - 1|, x^* \in S_A \\ N_\theta(A) + |\Pi_\theta(A) - \frac{r}{M}|, x^* \notin S_A \end{cases}$$

Given a training set containing $n$ pairs $(\tilde{\boldsymbol{x}}_i, x_i^*)$, the final empirical score is: $S(\theta) = \frac{1}{n}\sum_{i=1}^{n} S_i(\theta)$ and training consists of finding the $\theta$ that minimizes $S$.

## B.2 Application to the imperfect Lorenz 96 system

To test our framework, we reproduce the experiment designed by Williams et al. [2014], who used an imperfect L96 model [Lorenz, 1996] to generate ensemble predictions and investigate the performance of ensemble post-processing methods for the prediction of extreme events. The training sets consist of $4000$ independent pairs of EPS of size $M = 12$ and the associated observations, for each lead time $\tau = \{1, 3, 5, 7\}$ days[3]. The EPS have beforehand been pre-processed to remove the constant bias. The testing set consists of another $10,000$ independent pairs of bias-corrected EPS and associated observations, for each lead time. We consider the prediction of an extreme event: $A_e = \{x \leq q_{0.05}\}$, where $q_{0.05}$ is the 0.05 quantile of the climatic distribution of $x$ and a common event $A_c = \{q_{0.5} \leq x \leq q_{0.6}\}$ . Results are compared against those given by a probabilistic post-processing, namely a Gaussian ensemble dressing.

We first assess the performance of each interpretation in terms of the empirical ignorance score relative to the climatology:

$$S_n(p_\theta, c) = \frac{1}{n}\sum_{i=1}^{n}\Big(IGN(r_\theta, x_i^*) - IGN(c, x_i^*)\Big) = -\frac{1}{n}\sum_{i=1}^{n}\log_2\left(\frac{r_\theta(x_i^*)}{c(x_i^*)}\right) \qquad \text{(Equation 3)}$$

where, following the work of Bröcker and Smith [2008], in the probabilistic framework, the predictive probability $p_\theta(x^*|\tilde{\boldsymbol{x}})$ is blended with the climatology $c(x^*)$ of the verification $x^*$: $r_\theta(x^*) = \alpha p_\theta(x^*) + (1 - \alpha)c(x^*)$. Our possibilistic framework is a mapping $\mathbb{R}^M \mapsto [0, 1] \times [0, 1]$, while the ignorance applies to a probabilistic prediction $\mathbb{R}^M \mapsto [0, 1]$. We consequently need to find a mapping from the dual measures $N$ and $\Pi$ to an equivalent probability. Since possibility and necessity measures can be seen as upper and lower bounds of a consistent probability measure, we can write $P(A) = \alpha N(A) + (1 - \alpha)\Pi(A)$ with $\alpha \in [0, 1]$ for any event $A$ of interest. Varying $\alpha$

[3] $\tau = 1$ corresponds to 0.2 model time units after initialization and can be associated with approximately 1 day in the real world [Lorenz, 1996].

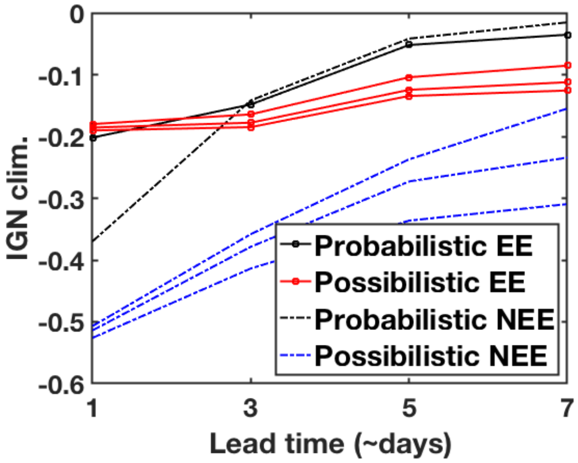


**Figure B.1** Ignorance relative to the climatology computed for the possibilistic (colored lines) and probabilistic (black lines) frameworks, in the case of the prediction of an extreme (EE; solid line) and a common (NEE; dashed line) event of interest, as defined in Sec. B.2. The upper and lower bounds, as well as the median, obtained by considering that $N(A) \leq P(A) \leq \Pi(A)$ in the possibilistic framework are reported.

allows one to browse across the range of associated probabilities, consistent with the possibility distribution $\pi$. We use this technique to compute the ignorance score of the possibilistic framework and compare its range to the performance of a probabilistic Gaussian ensemble dressing. Both frameworks are characterized by negative relative ignorance, confirming that they have a predictive added-value over climatology. The difference in ignorance equals the difference in expected returns that one would get by placing bets proportional to their probabilistic forecasts.

As shown in Figure B.1, for both types of events, the possibilistic framework performs as well or slightly better than the probabilistic, for all $\alpha \in [0, 1]$. The slight increase in performance remains relatively constant or even improve (extreme event case) with lead time. The relative ignorance of the possibilistic framework has a variance (due to the range of $\alpha$) that grows with the lead time, as expected.

To understand better the operational consequences of such results, we report in Figure B.2 the relative operating characteristic (ROC) of both frameworks at lead times of 3 and 7 days. Given a binary prediction (yes/no w.r.t. event $A$), the ROC plots the hit rate (HR; fraction of correctly predicted $A$ over all $A$ observed) versus the false alarm rate (FA; fraction of wrongly predicted $A$ over all $\bar{A}$ observed). We use increasing thresholds $p_t \in [0, 1]$ for making the decision (yes if $P(A) \geq p_t$) and report the associated HR and FA in the graph. Again, we vary $\alpha$ to see the range of HR and FA covered for each $p_t$ by the possibilistic prediction $(N, \Pi)$. The resulting points form a curve (probabilistic approach) or a cloud (possibilistic method), which are a visual way to assess the ability of a forecast system to discriminate between events and non events.

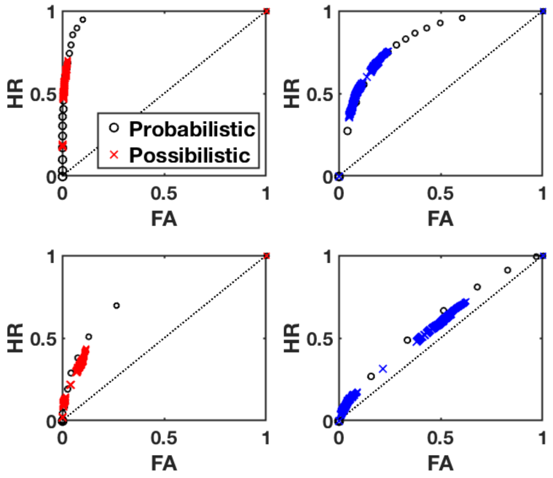


**Figure B.2** ROC curves for the extreme event (left side) and common event (right side) at lead time 3 days (top) and 7 days (bottom). The probabilistic results are reported by means of black circles and the possibilistic results by means of colored crosses. The larger the symbol, the larger the threshold probability used to compute HR and FA.

The possibilistic curves all fit or are very close to the probabilistic curves, for both extreme and common events and for all lead times. The main difference is their extension: the possibilistic framework remains located in areas of relatively small FA, compared to the results of the probabilistic approach for similar thresholds $p_t$. This results indicates that the HR remains smaller than what can be achieved by the probabilistic framework, showing lower skill. The fact that the possibilistic curves yet lies on the probabilistic ROC curves shows that the reason behind this discrepancy is not a lack of discrimination between events and non-events; for a given FA, both methods provide the same HR. The reason is connected to a bias in probabilities for the possibilistic approach towards zero and towards 1: the possibilistic framework is very sharp, as shown on the diagrams in Figure B.3. Because they are not blended with climatology, a large part of the predictions have zero probability associated to the event of interest, instead of a minimal one, which prevents the current implementation of the possibilistic framework from reaching higher HR. Side experimentation not reproduced here has shown that weighting the scores attributed to observed event $A$ in the global empirical training score allows to reproduce fully the probabilistic curve for each lead time.

Reliability diagrams presented on Figure B.4 plot the observed conditional frequencies against the corresponding forecast probabilities for lead time 3 and 7 days. They illustrate how well the predicted probabilities of an event correspond to their observed conditional frequencies. The predictive model is all the more reliable (i.e. actionable) when the associated curve is close to the diagonal. Noting that the diagonal represents perfect reliability, the distance to the diagonal indicates underforecasting

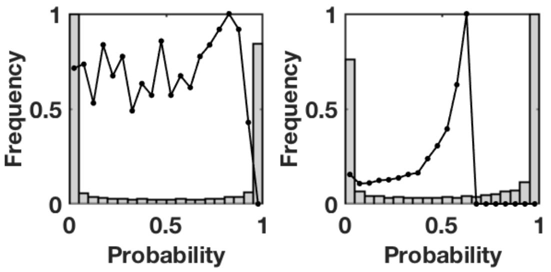


**Figure B.3** Normalized histograms of the equivalent forecast probabilities in the possibilistic framework for the observations of the extreme event (left) and common event (right) at lead time 3 days. The corresponding distributions of predictive probabilities in the probabilistic framework come on top as a thick black lines.

(curves above) or overforecasting (curves below). Distance above the horizontal climatology line indicates a system with resolution, a system that does discriminate 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.

The probabilistic curves are globally aligned with the perfect reliability line, yet with growing lead time, they are restricted to small probabilities only (because of wider EPS or pure predictability issues such as mentioned for extreme events). On the contrary, the reliability plots associated with the possibilistic approach cover all range of probabilities. This approach tends to be underforecasting (resp. overforecasting) for small (resp. large) probabilities, especially for the common event. A large part of the area covered by the possibilistic solutions is contained in the skill cones for the rare event, denoting a skilled predictive system for all but very low predictive probabilities. Results are less interesting for the common event, where the possibilistic framework leads to a flatter diagram, indicating less resolution, especially with larger lead times.

## B.3 Conclusions

In this work, we have presented a possibilistic framework which allows us to interpret ensemble predictions without the notion of *member density*, or *additivity* that proved to be incoherent with the conditions in which EPS were built. Preliminary results show that such a framework can be used to reproduce the probabilistic performances (ROC curves, resolution) and even slightly improve some of them (ignorance, sharpness, reliability). Moreover, the proposed approach addresses some of the well-known limitations of the probabilistic framework (reliability, for example). The added-value of this framework is particularly tangible for extreme events. Further work is needed to improve the design of the possibilistic distributions, by means of dynamical information or statistical priors. Besides, developments regarding the understanding and the operational use of such 'fuzzy' results are necessary.

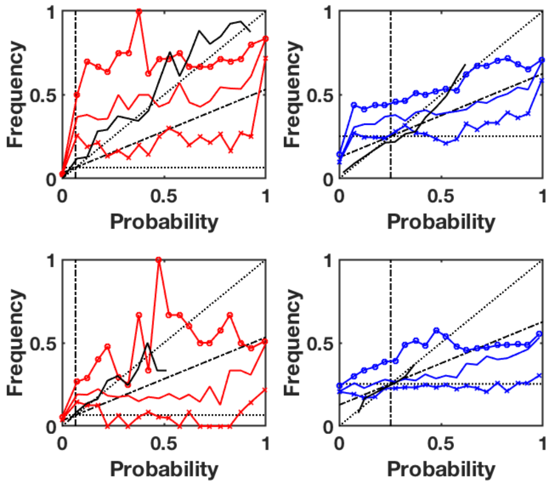


**Figure B.4** Reliability diagrams for the extreme event (left side) and common event (right side) at lead time 3 (top) and 7 days (bottom). The probabilistic results are reported in black line, while the upper, median and lower bounds of the possibilistic ones are in thinner red lines. Standards elements of comparison are reported in the diagram, as described in Sec. B.2, namely the diagonal (perfect reliability), the climatological reference (horizontal dotted) and the cones of skill (inside the dashed-dotted secants).

## Bibliography

Jochen Bröcker and Leonard A. Smith. From ensemble forecasts to predictive distribution functions. *Tellus A: Dynamic Meteorology and Oceanography*, 60(4):663–678, 2008. doi: 10.1111/j.1600-0870.2007.00333.x.

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

Edward N Lorenz. Predictability: A problem partly solved. In *Proc. Seminar on predictability*, volume 1, 1996.

Mark S. Roulston and Leonard A. Smith. Evaluating Probabilistic Forecasts Using Information Theory. *Monthly Weather Review*, 130(6):1653–1660, 2002. doi: 10.1175/1520-0493(2002)130<1653:EPFUIT>2.0.CO;2.

R. M. Williams, C. A. T. Ferro, and F. Kwasniok. A comparison of ensemble post-processing methods for extreme events. *Quarterly Journal of the Royal Meteorological Society*, 140(680):1112–1120, 2014. doi: 10.1002/qj.2198.