Post-traitement statistique des prévisions d'ensemble : théorie, application en météorologie et vérification

Romain Pic

UBFC
UNIVERSITÉ
BOURGOGNE FRANCHE-COMTÉ

cnrs

laboratoiredemathématiquesdebesançon
UFC • UBFC • CNRS • UMR 6623

Statistical postprocessing of ensemble forecasts : theory, application in weather forecasting and verification

Post-traitement statistique des prévisions d'ensemble : théorie, application en météorologie et vérification

Thèse de doctorat de l'Université Bourgogne Franche-Comté

préparée à l'Université de Franche-Comté

École doctorale Carnot Pasteur

présentée et soutenue publiquement à Besançon le 9 septembre 2024
en vue de l'obtention du grade de

Docteur de l'Université Bourgogne Franche-Comté

mention Mathématiques

par

Romain Pic

Composition du jury :

Mme. Chavez Valérie	Professeure, Université de Lausanne	Présidente
Mme. Thorarinsdottir Thordis	*Associate Professor*, University of Oslo	Rapportrice
M. Vannitsem Stéphane	*Head of Service*, Institut Royal Météorologique de Belgique	Rapporteur
Mme. Goga Camelia	Professeure des universités, Université de Franche-Comté	Examinatrice
M. Goude Yannig	*Senior researcher*, EDF R&D	Examinateur
M. Dombry Clément	Professeur des universités, Université de Franche-Comté	Directeur de thèse
M. Naveau Philippe	Directeur de recherche, CNRS	Co-directeur de thèse
M. Taillardat Maxime	Ingénieur des Travaux de la Météorologie, Météo-France	Co-directeur de thèse

Remerciements

Je souhaite commencer par grandement remercier mes directeurs de thèse, Clément, Philippe et Maxime. Nos échanges autour des nombreux sujets qui composent cette thèse ont été très enrichissants pour moi et continueront d'influencer mon approche de la recherche pendant un certain temps. J'ai particulièrement apprécié que mon avis ait été pris en compte de manière égale aux vôtres dans les différents projets de ma thèse, et ce, dès le début de mon stage. En dehors de la recherche, j'ai également apprécié les moments passés à vos côtés, à Aussois notamment.

Un immense merci à Thordis Thorarinsdottir et Stéphane Vannitsem d'avoir accepté de rapporter cette thèse. Je leur suis d'autant plus reconnaissant qu'ils ont pris de leur temps durant la période estivale, et plusieurs de leurs contributions respectives ont grandement aidé à ma compréhension de thématiques centrales à cette thèse. *Many thanks to Thordis Thorarinsdottir and Stéphane Vannitsem for agreeing to report on this thesis. I am especially grateful to them for taking time out of their summer schedules, and many of their respective contributions have greatly helped my understanding of themes central to this thesis.*

Merci également à Valérie Chavez, Camelia Goga et Yannig Goude d'avoir accepté de faire partie du jury de cette thèse. Je remercie tout particulièrement Camelia pour sa sympathie tout au long de ma thèse et notamment lors du cours que j'ai eu le plaisir de partager avec elle.

J'adresse mes remerciements aux membres du Laboratoire de Mathématiques de Besançon. Plus particulièrement, les membres de l'équipe Probabilités et Statistiques ainsi que mes collègues doctorants. Merci à Mathilde, Cécile, Mehdi et Valentin pour les moments que nous avons pu partager.

Toujours au sein du LmB, je tiens à remercier chaleureusement Charlène, Pascaline et Julien qui m'ont aidé à surmonter de nombreuses galères administratives et informatiques, et m'ont soutenu en me sortant de la solitude souvent liée à ces épreuves.

Je tiens à remercier les membres du Laboratoire des Sciences du Climat et de l'Environnement pour l'accueil qu'ils m'ont réservé durant mon séjour. Je suis également reconnaissant envers les équipes de Météo-France et du Centre National de Recherches Météorologiques pour leur soutien constant lors de mon utilisation de leurs ressources informatiques.

En plus de mes trois directeurs de thèse, je suis extrêmement reconnaissant envers Zeina Al Masry, Christine Devalland, Thibault Modeste et Julie Bessac pour l'opportunité que j'ai eue de travailler à leurs côtés sur des thématiques proches de mon sujet de thèse.

Je suis profondément reconnaissant envers les personnes que j'ai rencontrées lors des nombreuses conférences auxquelles j'ai eu l'occasion d'assister, et je suis impatient de poursuivre nos échanges à l'avenir.

Je voudrais remercier les personnes qui m'ont accompagné lors de ma (longue) formation. Merci à Mme Jariel, Mme Berger et M. Coulet qui m'ont aidé à cultiver mon amour pour les sciences notamment la physique, les mathématiques et l'informatique. Merci à José, Emmanuel,

Vincent et Irmgard de m'avoir initié à la recherche. Merci à Juliette, Vincent P., Mathilde, Toscan, Vincent D., Arthur et Ersin, mes compagnons de galère en prépa, en école et en stage d'avoir rendu le trajet moins long et plus agréable.

Je remercie du plus profond de mon cœur ma famille et mes amis, pour leurs encouragements. Merci Maman, Papa et Thomas de m'avoir permis de suivre cette voie et de m'avoir toujours soutenu.

Je souhaite remercier mes amis du Sud (a.k.a. les boss du lycée) pour leur présence tout au long de cette aventure qui a conduit à la réalisation de cette thèse. Merci à Arthur pour le soutien et les rires, peu importe l'heure de la journée et surtout de la nuit.

Merci, Mahaut, de m'avoir écouté parler de sujets incompréhensibles et parfois utiliser des métaphores équestres qui compliquaient souvent (toujours) le tout. Merci pour ton soutien au quotidien et tout au long de cette thèse.

Contents

Chapter 1

Introduction

Accurate weather prediction is crucial in various fields (e.g., renewable energy, transport networks or farming) for both decision-making and its financial impact (Palmer, 2002). Probabilistic forecasts are an essential component of optimal decision-making as they quantify the uncertainty of the prediction (Gneiting and Katzfuss, 2014). In weather forecasting, statistical postprocessing is necessary to produce calibrated and sharp probabilistic forecasts from ensemble prediction systems. This thesis focuses on three different aspects of statistical postprocessing: theoretical convergence rates, grid-based postprocessing of precipitation and verification of spatial probabilistic forecasts.

This thesis was conducted in collaboration with Météo-France via the direct supervision of Maxime Taillardat but also through exchanges with other members of Météo-France and the use of their high-performance computing resources.

Statistical postprocessing methods use the output of a physical model to improve the prediction of a variable of interest. Scoring rules are used for probabilistic forecast verification to measure and compare the predictive performance of competing forecasts. This thesis studies different aspects of statistical postprocessing and verification of probabilistic forecasts.

- From a theoretical point of view, only limited results are available regarding the convergence of postprocessing methods. Chapter 2 (Pic et al., 2023) is a theoretical contribution that focuses on the optimal minimax rate of convergence for the theoretical risk associated with the continuous ranked probability score.

- In weather forecast applications, spatial forecasts are ubiquitous. However, the random forest-based methods that are used operationally to postprocess forecasts at Météo-France do not really take into account the spatial setting. Furthermore, they suffer from storage memory voracity and an inability to extrapolate. In Chapter 3 (Pic et al., 2024b), we propose a U-Net-based distributional regression method to postprocess ensembles circumventing storage memory voracity while achieving a predictive performance comparable to state-of-the-art methods.

- Regarding probabilistic forecast verification, no single scoring rule is able to provide an ideal assessment of the predictive performance of forecasts, and thus, different scoring rules should be used to understand it. This statement is even more important in a spatial forecast verification setting as predictive performance is subject to complex characteristics. With that in mind, interpretable scoring rules are powerful tools facilitating forecast verification. Chapter 4 (Pic et al., 2024a) presents how aggregation and transformation principles can be used to construct interpretable proper multivariate scoring rules.

In addition, Chapter 5 provides perspectives on the works composing this PhD thesis.

The remainder of this chapter is organized as follows. Section 1.1 introduces the context necessary to the comprehension of the contribution of this thesis. Section 1.2, Section 1.3 and Section 1.4 summarize works related in Chapter 2, Chapter 3 and Chapter 4, respectively. Section 1.5 briefly summarizes the works related in Appendix A and Appendix B.

1.1 General introduction

1.1.1 Uncertainty in deterministic systems modeling and ensemble forecasts

An intuitive approach to weather prediction is to consider that the physics of the atmosphere is governed by a set of deterministic nonlinear differential equations (Bjerknes, 2009). However, in the 1960s, Lorenz (1963) has shown that the atmosphere is a chaotic system characterized by multiple sources of uncertainty (see Wilks and Vannitsem 2018 for more details).

Sensitivity to initial conditions combined with initial conditions uncertainty represents a major source of uncertainty for weather prediction. Initial conditions uncertainty arises from various aspects, such as the combination of different types of observations and the varying quality and coverage depending on the variable of interest, location, and mean of measurement. For example, as used in Chapter 3, radar-based precipitation measurement quality depends on the distance to the instrument and the underlying orography (see, e.g., Germann et al. 2022). The field of data assimilation is dedicated to the combination of different sources of data to provide well-suited initial conditions for numerical weather prediction (NWP) systems. Additionally, in practice dynamical weather forecast models do not perfectly describe the true dynamics. First, the model might provide an incorrect modelization of the phenomena at play. Second, all NWP systems are incomplete due to spatial and temporal discretization and the parametrization of unresolved physical processes.

Moreover, the atmosphere has a flow-dependent predictability, meaning that the propagation of initial condition uncertainty depends on the state of the system. This causes forecast errors to fluctuate across the globe and depending on the variable of interest but also from one day to another (Buizza, 2018). These limitations also affect other physical models such as climate models and hydrology models.

Ensemble forecasts have been developed as a means to try to capture model uncertainties. However, choosing a well-suited ensemble is challenging, as random sampling based on a range of possible outcomes does not lead to an informative ensemble. Moreover, a large number of members may be appealing but computationally expensive, and an increase in resolution is often preferred as it enables resolving processes at finer scales. In order to sample a system with millions of degrees of freedom with few tens of members, different approaches have proven to be capable of representing model uncertainties: multimodel approaches, perturbed approaches, perturbed-tendency approaches, stochastic back-scatter approaches, and combinations of them. Readers may refer to Buizza (2018) for a historical overview of the use of ensemble forecasting.

Regardless of the continuous improvement of NWP systems over the past decades (Bauer et al., 2015), the improvement of near-surface variables' predictive performance is slower than that of variables higher in the atmosphere (Buizza, 2018). Ensemble forecasts issued from NWP systems suffer from bias and underdispersion. This phenomenon affects all NWP systems regardless of the weather service and of the physical variable of interest. As the dynamical systems they rely on, ensemble forecast errors vary depending on the variable of interest and the region of interest. Moreover, the increase in lead time (i.e., time between when a forecast is issued and its validity) is associated with a decrease in predictability. As these errors are systematic they can be corrected by statistical approaches called *statistical postprocessing methods*.

1.1.2 Statistical postprocessing

Statistical postprocessing methods aim to use previous pairs of the raw ensemble and observations to improve prediction of a variable of interest. The *raw ensemble* denotes the unprocessed (i.e., raw) ensemble output of NWP systems. As the aim is to provide forecasts that are informative to end-users, forecasts should be probabilistic forecasts. *Probabilistic forecasts* provide a prediction in the form of a probability distribution. This enables the quantification of the prediction uncertainty, ensuring optimal decision-making (Gneiting and Katzfuss, 2014). The raw ensemble is a probabilistic forecast as it can be interpreted as an empirical distribution where all members have the same probability. Probabilistic forecasts can take the form of any formulation able to describe the whole probability distribution. In a univariate setting, they can take the form of a probability density function, a cumulative distribution function, or a quantile function, for example.

Statistical postprocessing methods can be classified in multiple manners. We present three different classifications of methods based on their distributional parametrization, their usage and their complexity. First, statistical postprocessing methods can be classified into two groups ("nonparametric" and "parametric") based on the assumption of a family of distributions. Nonparametric methods include analog ensemble (see, e.g., Delle Monache et al. 2013) which uses similar past atmospheric situations to improve on the raw ensemble. Analog ensemble is related to k-nearest neighbor (k-NN) methods as explained in Appendix C. Quantile regression forest (QRF; Taillardat et al. 2016) is a nonparametric method using the data in terminal nodes (i.e., leaves) of a random forest to compute a weighted average of empirical distributions. Parametric methods include ensemble model output statistic (EMOS; Gneiting et al. 2005) assuming that the predicted distribution is a normal distribution with parameters linearly depending on the summary statistics of the raw ensemble. Parametric methods provide a family of parametric distributions suited to the variable of interest (e.g., based on extreme value theory; Friederichs et al. 2018). Most nonparametric methods lack any extrapolation availability beyond the range of observed data but are able to conserve the characteristics of the true distribution from the observed data. The frontier between the two classes is porous: QRF with tail extension (TQRF; Taillardat et al. 2019) is a semi-parametric method fitting a parametric distribution on the output of a QRF. The classification into parametric and nonparametric methods is discussed in greater detail in Vannitsem et al. (2021).

Second, statistical postprocessing methods also differ in their usage. The most common usage of postprocessing is to separately postprocess univariate marginals and the dependence structure. The dependence structure can be obtained from the raw ensemble as done by ensemble copula coupling (ECC; Schefzik et al. 2013) or from historical observations as done by Schaake shuffle (ScS; Clark et al. 2004). Alternatively, if the raw ensemble or historical data do not model the dependence structure sufficiently well, it can be postprocessed using adapted techniques such as a Gaussian copula approach (see, e.g., Möller et al. 2013). Certain statistical postprocessing methods directly consider multivariate quantities (e.g., Pinson et al. 2009). Readers may refer to Schefzik and Möller (2018) for more details on dependence structure postprocessing. Some methods directly postprocess each member of the raw ensemble simultaneously. Member-by-member (Van Schaeybroeck and Vannitsem, 2015) correct for ensemble mean and spread via a linear combination of the raw predictors. Postprocessing of ensembles with transformers (PoET; Ben Bouallègue et al. 2024b) uses transformers within a U-Net architecture to postprocess ensemble members.

Third, statistical postprocessing methods differ in their level of complexity. Less complex methods are related to statistical learning methods as for analog ensemble with k-NN and QRF with random forests (Breiman, 2001; Meinshausen, 2006). The relative simplicity of these methods allows for more simplicity but less flexibility in terms of modeling the dependence of a variable of interest given predictors. More complex methods issued from machine learning

can also be employed. Distributional regression networks (DRN; Rasp and Lerch 2018) is a neural network (NN-)based approach predicting the parameters of a distribution of interest. It leverages the flexibility of fully connected NN to model the dependency of the parameters on the covariables (used as input of DRN). DRN can be seen as an extension of EMOS. Instead of linearly modeling the dependence of the parameters on summary statistics of the raw ensemble, it allows for more flexible nonlinear dependencies to be accounted for. At the upper end of the complexity spectrum are methods based on deep learning (DL) techniques. The PoET approach, introduced above, uses transformers that were originally introduced for natural language processing tasks (Vaswani et al., 2017). Not all DL-based methods represent the same level of complexity. Complexity can be accompanied by an increase in flexibility but also in terms of difficulty of implementation. Chapter 3 proposes a U-Net-based method to predict parametric distributions which extends DRN to grid-based data.

The three classifications provide a first view of the wide scope of statistical postprocessing methods. For more detailed overviews, readers may refer to Taillardat et al. (2019), Vannitsem et al. (2021) and Schulz and Lerch (2022b).

As briefly mentioned, depending on the application, different postprocessing methods could be preferred. We briefly focus on parametric methods to explicitly discuss how variables of interest differ in terms of postprocessing. First, different families of distribution are suited to different variables. For example, temperature and sea-level pressure can be modeled by normal distributions. Other variables may present asymmetric, multimodal or discontinuous distributions which can be modeled using truncated, censored or mixed distribution families. For example, rainfall presents an atom mass in zero related to dry events (i.e., absence of rainfall). Moreover, rainfall is often heavy-tailed; thus, a distribution family leveraging extreme-value theory can improve its postprocessing. Lerch and Thorarinsdottir (2013) proposed a variant of EMOS using a generalized extreme-value distribution to postprocess maximum daily wind speed. Taillardat et al. (2019) introduces TQRF as an extension of QRF to improve the prediction of extreme rainfall. A complete review of postprocessing for extreme events is provided in Friederichs et al. (2018).

Moreover, Hemri et al. (2014) and Taillardat and Mestre (2020) have highlighted that all variables of interest do not represent the same difficulty in terms of postprocessing. Variables with short-scale spatio-temporal dependence (e.g., rainfall or wind gusts) are more difficult to treat than spatially smooth variables (e.g., surface temperature or sea-level pressure). In the same vein, Schulz and Lerch (2022b) states that "wind gusts are a challenging meteorological target variable as they are driven by small-scale processes and local occurrence, so that their predictability is limited even for numerical weather prediction (NWP) models run at convection-permitting resolutions." The predictability of variables is related to their physical characteristics and to their representation within NWP models.

1.1.3 Verification of probabilistic forecasts

Verification of probabilistic forecasts fulfills two main purposes: quantifying a forecast's predictive performance and comparing competing forecasts. In the context of statistical postprocessing, the obvious forecast of reference is the raw ensemble and postprocessing techniques should improve predictive performance with respect to this baseline.

Gneiting et al. (2007) proposed a paradigm for the evaluation of probabilistic forecasts: "maximizing the sharpness of the predictive distributions subject to calibration." *Calibration* is the statistical compatibility between the forecast and the observations. *Sharpness* is the concentration of the forecast and is a property of the forecast itself. In other words, the paradigm aims at minimizing the uncertainty of the forecast given that the forecast is statistically consistent with the observations. This principle for the evaluation of probabilistic forecasts has reached a consensus in the field of probabilistic forecasting (see, e.g., Gneiting and Katzfuss

2014; Thorarinsdottir and Schuhen 2018).

For univariate forecasts, multiple definitions of calibration are available depending on the setting. The most used definition is *probabilistic calibration* and, broadly speaking, it consists of computing the rank of observations among samples of the forecast and checking for uniformity with respect to observations. If the forecast is calibrated, observations should not be distinguishable from forecast samples, and thus, the distribution of their ranks should be uniform. Probabilistic calibration can be assessed by probability integral transform histograms (Dawid, 1984) or rank histograms (Anderson, 1996; Talagrand et al., 1997) for ensemble forecasts when observations are stationary (i.e., their distribution is the same across time). Readers interested in a more in-depth understanding of univariate forecast calibration are encouraged to consult Tsyplakov (2013, 2020). For multivariate forecasts, a popular approach relies on a similar principle: first, multivariate forecast samples are transformed into univariate quantities using so-called pre-rank functions and then the calibration is assessed by techniques used in the univariate case (see, e.g., Gneiting et al. 2008; Allen et al. 2024).

With a quantitative perspective, scoring rules provide a quantitative assessment of the quality of a probabilistic forecast in view of the observation that materializes. A scoring rule S assigns a real-valued quantity $\mathrm{S}(F, y)$ to a forecast-observation pair (F, y), where $F \in \mathcal{F}$ is a probabilistic forecast and $\boldsymbol{y} \in \mathbb{R}^m$ is an observation. In the negative-oriented convention, a scoring rule S is *proper relative to the class* $\mathcal{F}$ if

$$\mathbb{E}_G[\mathrm{S}(G, \boldsymbol{Y})] \leq \mathbb{E}_G[\mathrm{S}(F, \boldsymbol{Y})] \tag{1.1}$$

for all $F, G \in \mathcal{F}$, where $\mathbb{E}_G[\cdots]$ is the expectation with respect to $\boldsymbol{Y} \sim G$. In simple terms, a scoring rule is proper relative to a class of distribution if its expected value is minimal when the true distribution is predicted, for any distribution within the class. Moreover, the scoring rule S is *strictly proper relative to the class* $\mathcal{F}$ if the equality in (1.1) holds if and only if $F = G$. This ensures the characterization of the ideal forecast (i.e., there is a unique forecast associated with the minimal expectation and it is the true distribution). Moreover, proper scoring rules are powerful tools as they allow the assessment of calibration and sharpness simultaneously (Winkler, 1977; Winkler et al., 1996).

However, as recalled in Chapter 4, (strict) propriety solely is not sufficient to lead to informative scoring rules. We propose a framework to construct interpretable proper scoring rules that are more informative in the verification of spatial probabilistic forecasts.

Moreover, since statistical postprocessing methods learn to predict a probabilistic distribution based on past observations, their practical evaluation should be based on an independent set of unseen data to avoid potential bias. In practice, additional limitations may arise from seasonality or the lack of data consistency (e.g., due to climate change or NWP systems updates).

1.2 Distributional regression and its evaluation with the CRPS: bounds and convergence of the minimax risk

Numerous statistical postprocessing methods rely on distributional regression. Postprocessing methods aim to model the conditional distribution of a variable of interest $Y \in \mathbb{R}^m$ (e.g., 3-h accumulated precipitation) given the output of a physical model $X \in \mathbb{R}^d$ (e.g., in the form of summary statistics), denoted F^*_X. In a verification setting, scoring rules are used to measure and compare the predictive performance of competing probabilistic forecasts. Scoring rules can be seen as the equivalent of loss functions (also known as scoring functions) in point regression.

Let $\overline{\mathrm{S}}(F, G) = \mathbb{E}_G[\mathrm{S}(F, Y)]$ denotes the expected score of F for the scoring rule S. In distributional regression, the predictive performance of a probabilistic forecast $\hat{F} : x \mapsto \hat{F}_x$ is assessed by its theoretical risk

$$\begin{aligned} R_P(\hat{F}) &= \mathbb{E}_{(X,Y)\sim P}\left[\mathrm{S}(\hat{F}_X, Y)\right]; \\ &= \mathbb{E}_{X\sim P_X}\left[\overline{\mathrm{S}}(\hat{F}_X, F^*_X)\right], \end{aligned}$$

where P is the joint distribution of (X, Y) and P_X is the marginal distribution of X. If S is strictly proper, then F^* is a Bayes predictor and its theoretical risk

$$\begin{aligned} R_P(F^*) &= \mathbb{E}_{(X,Y)\sim P}[\mathrm{S}(F^*_X, Y)]; \\ &= \mathbb{E}_{X\sim P_X}\left[\overline{\mathrm{S}}(F^*_X, F^*_X)\right] \end{aligned}$$

is the Bayes risk. We recall that the Bayes risk is the minimal theoretical risk over all possible predictors and that a Bayes predictor is a predictor achieving the Bayes risk. Moreover, if S is strictly proper, the set of Bayes predictors are the forecasts $\hat{F}$ such that $\hat{F}_X = F^*_X$ P_X-almost everywhere.

Statistical postprocessing techniques rely on a training sample $D_n = \{(X_i, Y_i), 1 \leq i \leq n\}$ and are evaluated in terms of their predictive performance with respect to new data (X, Y). Both the training sample D_n and the test data (X, Y) are independent and identically distributed from the same distribution P. Given the training sample D_n, an algorithm $\hat{F}_n : x \mapsto \hat{F}_{n,x}$ is constructed to estimate the conditional distribution F^*_x. In this context, the theoretical risk of $\hat{F}_n$ is expressed as

$$\begin{aligned} \mathbb{E}_{D_n\sim P^n}\left[R_P(\hat{F}_n)\right] &= \mathbb{E}_{D_n\sim P^n}\mathbb{E}_{(X,Y)\sim P}\left[\mathrm{S}(\hat{F}_{n,X}, Y)\right]; \\ &= \mathbb{E}_{D_n\sim P^n}\mathbb{E}_{X\sim P_X}\left[\overline{\mathrm{S}}(\hat{F}_{n,X}, F^*_X)\right]. \end{aligned}$$

The theoretical risk is averaged over possible values of the training sample D_n, making it solely dependent on the distribution P and the sample size n. As previously, if the scoring rule S is proper, F^* is a Bayes predictor and its risk is the Bayes risk. Then, the quantity of interest is the excess of risk defined as the difference between the theoretical risk of an algorithm $\hat{F}_n$ and the Bayes risk :

$$\mathbb{E}_{D_n\sim P^n}\left[R_P(\hat{F}_n)\right] - R_P(F^*) = \mathbb{E}_{D_n\sim P^n}\mathbb{E}_{X\sim P_X}\left[\overline{\mathrm{S}}(\hat{F}_{n,X}, F^*_X) - \overline{\mathrm{S}}(F^*_X, F^*_X)\right]. \tag{1.2}$$

When S is proper, the difference of expected scores within the expectation on the right-hand side is called the divergence of S (see, e.g., Gneiting and Katzfuss 2014, Section 3.1 and Thorarinsdottir et al. 2013).

We are interested in convergence results in distributional regression. The framework introduced above is widely used in practice but lacks theoretical guarantees. Most convergence statements in distributional regression are not only derived within an unconditional framework but also assume arbitrarily large sample sizes (see, e.g., Thorey et al. 2017 and Mösching and Dümbgen 2020). One exception is the isotonic distributional regression, which, under monotonicity assumptions, was shown to minimize the in-sample continuous ranked probability score (CRPS) and to satisfy consistency in the sense of the Kolmogorov distance (Henzi et al., 2021).

We focus on the univariate case ($m = 1$) as it corresponds to the setting of numerous statistical postprocessing methods. Moreover, we choose the scoring rule of interest to be the CRPS (Matheson and Winkler, 1976) which benefits from being strictly proper relative to $\mathcal{P}_1(\mathbb{R})$ (i.e., distributions on $\mathbb{R}$ with a finite first moment). Since $m = 1$, distributions are identified to their cumulative distribution function (cdf). The divergence of the CRPS is the L^2-norm of the

difference between the cdf of $\hat{F}_X$ and the conditional cdf F^*_X (also known as the squared second-order Cramér's distance or the integrated quadratic distance; Thorarinsdottir et al. 2013).

In point regression, it is necessary to restrict the convergence over a given class of distribution to obtain non-trivial results (Stone 1982; Györfi et al. 2002). In order to study the rate of convergence of the excess risk (1.2) to zero as $n \to \infty$, we introduce the notion of *optimal minimax rate of convergence*. The minimax risk corresponds to the best achievable risk in the worst-case scenario (whence the name minimax). More precisely, given a class of distributions $\mathcal{D}$, the optimal minimax rate of convergence quantifies the minimal error that an algorithm $\hat{F}_n$ can achieve uniformly on a given class of distributions $\mathcal{D}$, when the size n of the training set D_n gets large. The formal definition of minimax rate of convergence for distributional regression is as follows.

Definition 1.1. *A sequence of positive numbers (a_n) is called an optimal minimax rate of convergence on the class $\mathcal{D}$ if*

$$\liminf_{n\to\infty} \inf_{\hat{F}_n} \sup_{P\in\mathcal{D}} \frac{\mathbb{E}_{D_n\sim P^n}[R_P(\hat{F}_n)] - R_P(F^*)}{a_n} > 0 \tag{1.3}$$

and

$$\limsup_{n\to\infty} \inf_{\hat{F}_n} \sup_{P\in\mathcal{D}} \frac{\mathbb{E}_{D_n\sim P^n}[R_P(\hat{F}_n)] - R_P(F^*)}{a_n} < \infty, \tag{1.4}$$

where the infimum is taken over all distributional regression models $\hat{F}_n$ trained on D_n. If the sequence (a_n) satisfies only the lower bound (1.3), it is called a lower minimax rate of convergence.

We consider the following classes of distributions.

Definition 1.2. *For $h \in (0,1]$, $C > 0$ and $M > 0$, let $\mathcal{D}^{(h,C,M)}$ be the class of distributions P such that $F^*_x(y) = P(Y \le y|X = x)$ satisfies:*

i) $X \in [0,1]^d$ P_X*-a.s.;*

ii) For all $x \in [0,1]^d$, $\int_{\mathbb{R}} F^*_x(z)(1 - F^*_x(z))\mathrm{d}z \le M$;

iii) $\|F^*_{x'} - F^*_x\|_{L^2} \le C\|x' - x\|^h$ *for all* $x, x' \in [0,1]^d$.

The conditions defining the class of distribution $\mathcal{D}^{(h,C,M)}$ are analogous to the conditions in point regression. We provide an interpretation and discussion of the conditions $i) - iii)$. Condition $i)$ is a condition on the covariables and can be extended to a compact. This condition arises from the fact that increasing the number of samples n tries to fill the covariates space in order to have a training sample representative of all the possible values. Hence, every covariable point needs to be approachable and the span of the covariable space impacts the convergence. Condition $ii)$ bound the sharpness (or entropy) of F^*_x for all $x \in [0,1]^d$. Sharpness is associated with the information carried by distribution, and it may appear intuitive that the less information is carried (i.e., the larger M), the more samples are required to obtain the same predictive performance (in terms of theoretical risk). Condition $iii)$ is a regularity condition imposing that close covariates lead to close conditional distributions. Since the algorithm $\hat{F}_n$ uses knowledge from past observations to estimate the conditional distribution at $X = x$ and increasing the number of sample n leads to having training data closer to $X = x$, regularity of F^* is needed to ensure that the increase in n is associated with an increase of predictability.

Definition 1.1 can be restated: an optimal minimax rate of convergence on the class $\mathcal{D}$ is a lower minimax rate of convergence on $\mathcal{D}$, and there exists an algorithm $\hat{F}_n$ achieving this rate.

We are able to obtain a lower minimax rate of convergence using a subclass of $\mathcal{D}^{(h,C,M)}$ which reduces the problem to standard results of point regression (Györfi et al., 2002).

In order to find an algorithm $\hat{F}_n$ such that it achieves the lower minimax convergence rate of convergence obtained, we investigate k-nearest neighbor (k-NN) methods and kernel methods. k-NN is well-known in the classical framework of regression and classification (see, e.g. Biau and Devroye 2015). In distributional regression, k-NN can be suitably adapted to estimate the conditional distribution F_x^* and the estimator is written as

$$\hat{F}_{n,x}(z) = \frac{1}{k_n} \sum_{i=1}^{k_n} \mathbb{1}_{Y_{i:n}(x) \leq z}, \tag{1.5}$$

where $1 \leq k_n \leq n$ and $Y_{i:n}(x)$ denotes the observation at the i-th nearest neighbor of x. As usual, possible ties are broken at random to define nearest neighbors.

The kernel estimate in distributional regression (see, e.g., Györfi et al. 2002, Chapter 5) can be expressed as

$$\hat{F}_{n,x}(z) = \frac{\sum_{i=1}^n K(\frac{x-X_i}{h_n}) \mathbb{1}_{Y_i \leq z}}{\sum_{i=1}^n K(\frac{x-X_i}{h_n})}, \tag{1.6}$$

if the denominator is nonzero. When the denominator is zero, we use the convention $\hat{F}_{n,x}(z) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{Y_i \leq z}$. Here the bandwidth $h_n > 0$ depends on the sample size n, and the function $K : \mathbb{R}^d \to [0, \infty)$ is called the kernel.

We obtain explicit and non-asymptotic upper bounds for the excess risk (1.2) of the k-NN and the kernel methods, respectively, uniformly on $\mathcal{D}^{(h,C,M)}$. Optimizing the bounds with respect to suitable choices of k_n and h_n leads to the following results on the optimal minimax rate of convergence.

Theorem 1.1. *The sequence $a_n = n^{-\frac{2h}{2h+d}}$ is the optimal minimax rate of convergence on the class $\mathcal{D}^{(h,C,M)}$.*

In particular, k-NN and kernel methods reach the optimal minimax rate of convergence for in dimension $d \geq 2$ and in any dimension d, respectively. In the context of statistical postprocessing, k-NN, and kernel methods are related to analog ensemble techniques (see, e.g., Delle Monache et al. 2013), and this relation is discussed in more detail in Appendix C. Additional comments on Chapter 2 are provided in Appendix C.

Chapter 2 : Summary of contributions

- We formalize a framework to adapt concepts of estimation theory to study theoretical risks in terms of scoring rules.
- We obtain the optimal minimax rate of convergence in distributional regression for a given class of distribution (Theorem 2.1).
- k-NN and kernel methods reach the optimal minimax rate of convergence in dimension $d \geq 2$ and in any dimension d, respectively.
- We obtain non-asymptotic upper bounds on the convergence rate for k-NN and kernel methods with a fixed sample size n (Propositions 2.1 and 2.2).
- The results can be extended to the threshold-weighted CRPS (see Appendix C).

1.3 Distributional regression U-Nets for the postprocessing of precipitation ensemble forecasts

Operationally at Météo-France, temperature and precipitation forecasts postprocessing rely on local (i.e., one model per location) random forest-based models (Taillardat and Mestre, 2020). Quantile regression forests (QRF; Meinshausen 2006) is a nonparametric method able to predict conditional quantiles or, more generally, a conditional distribution. Similarly to random forests, it uses the data in terminal nodes (i.e., leaves) to compute a weighted average of empirical distributions. QRFs have proven their performance for a large variety of variables (Taillardat et al., 2016; Whan and Schmeits, 2018; van Straaten et al., 2018; Rasp and Lerch, 2018; Taillardat et al., 2019; Schulz and Lerch, 2022b). QRFs are known to have three main limitations: potential spatial inconsistency, storage memory voracity (Taillardat and Mestre, 2020), and inability to extrapolate.

Rasp and Lerch (2018) proposed distributional regression networks (DRN), a neural network (NN-)based global model predicting the parameters of a distribution of interest. It leverages the flexibility of NN to model the dependency of the parameters on the covariables (used as input of DRN). DRN can be seen as an extension of EMOS (Gneiting et al., 2005), which itself fits a parametric distribution where the parameters linearly depend on summary statistics of the raw ensemble. DRN is a global model thanks to the presence of an embedding module within its architecture, allowing the network to learn location-specific parameters and to benefit from data at similar locations. Rasp and Lerch (2018) and Schulz and Lerch (2022b) have shown that DRN outperforms other state-of-the-art methods in most stations over Germany for the postprocessing of temperature and wind gusts, respectively. In spite of being a global model, the architecture of DRN makes it ill-suited to gridded data. Its architecture does not use knowledge of the spatial structure of the points and thus has to try to learn it through its embedding module.

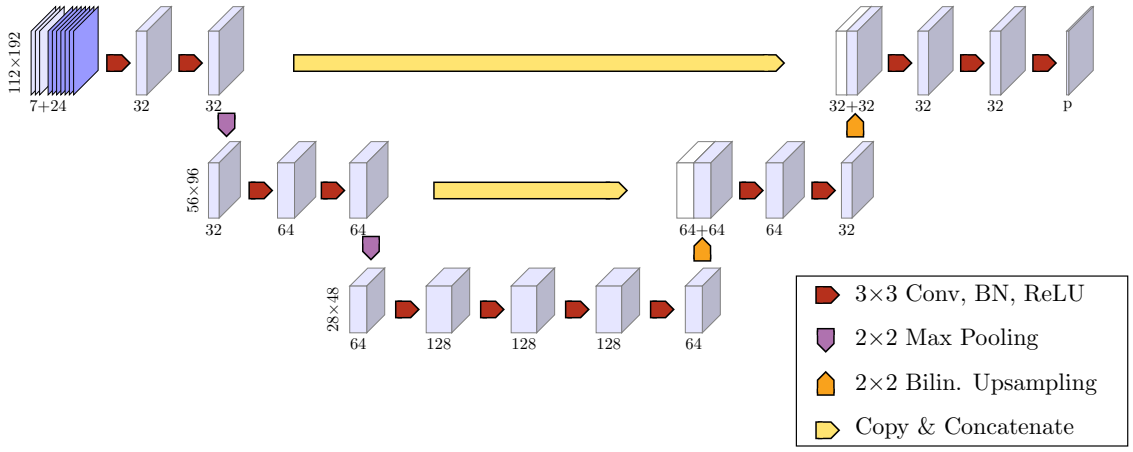


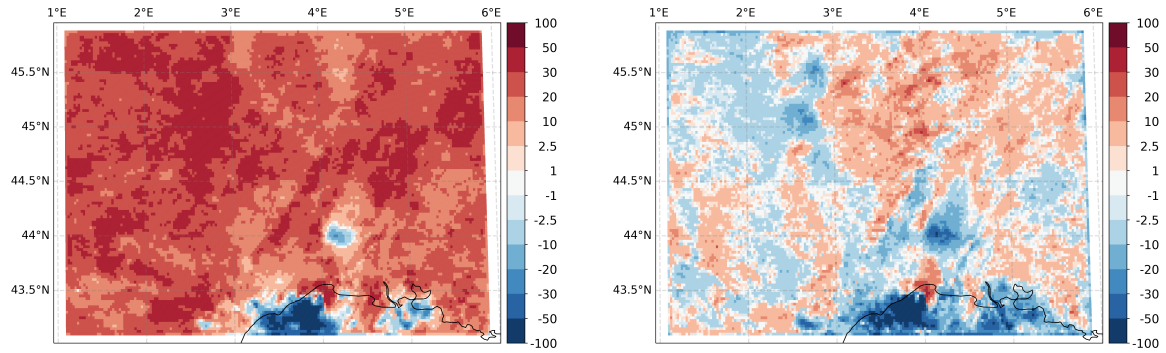
Figure 1.1: Architecture of distributional regression U-Nets. *Conv* stands for convolution, *BN* stands for batch normalization, *ReLU* stands for rectified linear unit and *Bilin. Upsampling* stands for bilinear upsampling. p is the number of distribution parameters: for GTCND and CSGD, $p = 3$.

We propose a U-Net-based distributional regression method suited to gridded data, which predicts parameters of marginal distributions at each grid point. Distributional regression U-Net (DRU) takes as input both constant fields (e.g., orography) and summary statistics of the raw ensemble (see Figure 1.1). On the left part, the succession of specific convolutional blocks

(red arrows and purple arrows) leads to an increase in the number of features and a reduction of the spatial dimension (i.e., a coarsening of the spatial resolution) as the data progresses through the network. These convolutional blocks are constructed in order to learn useful representations of the features of the fields at various spatial scales. On the right part, upscaling blocks (orange arrows), based on bilinear upsampling, use the features learned in the central part of the architecture to predict features at finer resolutions and finally learn the parameters of the distribution selected. Additionally, we use skip-connections (yellow arrows) since they have proven to improve the stability of the convergence of NN (Li et al., 2018).

We focus on 3-h accumulated precipitation over the South of France, which is a region prone to Mediterranean heavy precipitation events. Ensemble forecasts are taken from the 17-member ensemble forecasting system AROME-EPS (Bouttier et al., 2015) which produces a gridded ensemble over Western Europe with a horizontal resolution of 0.025°. Probabilistic forecasts are compared to 3-h accumulated precipitation data obtained from the gauge-adjusted radar product ANTILOPE (Champeaux et al., 2009). Precipitation is a challenging meteorological quantity to predict due to its heavy-tailed climatology and its short-scale spatio-temporal dependence (Hemri et al., 2014; Taillardat and Mestre, 2020). Moreover, another challenging aspect of the dataset is that only three years of training data are available. This is especially challenging regarding the prediction of high precipitation.

To suit the prediction of precipitation, two parametric distributions with an atom mass in zero (i.e., for dry events) are selected: the generalized truncated/censored normal distribution (GTCND; Jordan et al. 2019) and the censored-shifted gamma distribution (CSGD; Scheuerer and Hamill 2015a). We denote U-Net+*distrib* the DRU where *distrib* is the parametric distribution.



(a) CRPSS of U-Net+CSGD w.r.t. raw

(b) CRPSS of U-Net+CSGD w.r.t. QRF

Figure 1.2: Predictive performance of U-Net+CSGD in terms of CRPS. CRPSS w.r.t. (a) the raw ensemble and QRF (b) of U-Net+CSGD.

Chapter 3 provides a thorough comparison between the raw ensemble, QRF, QRF with tail extension (TQRF; Taillardat et al. 2019) and DRU. Here, we provide a simplified comparison between U-Net+CSGD, the raw ensemble and QRF only. In terms of CRPS, the relative improvement can be expressed using the continuous ranked probability skill score (CRPSS) defined as

$$\mathrm{CRPSS}(F, F_{\mathrm{ref}}) = 1 - \frac{\mathbb{E}_G[\mathrm{CRPS}(F, Y)]}{\mathbb{E}_G[\mathrm{CRPS}(F_{\mathrm{ref}}, Y)]},$$

where G is the distribution of the observations and $\mathbb{E}_G[\cdots]$ is the expectation with respect to $Y \sim G$.. The CRPSS is positive if the forecast F improves the expected CRPS w.r.t. the reference forecast F_{ref} and negative otherwise. In the following, the CRPSS is expressed in

percentage. Figure 1.2 provides the CRPSS of U-Net+CSGD w.r.t. raw ensemble and QRF. U-Net+CSGD leads to a CRPSS w.r.t. the raw ensemble of 22.36% when averaged over the region of interest. As QRF, DRUs lead to improvement in terms of CRPSS over the vast majority of grid points. Nonetheless, there are areas where they have a poorer predictive performance compared to raw ensemble. These areas are located over the Mediterranean Sea or near the coast, and one patch is located in the Rhône River valley. This is caused by the fact that the area over the Mediterranean Sea is associated with the lowest precipitation accumulations and lower observation quality since it is far from the nearest radar and cannot be corrected by gauges. Overall, U-Net+CSGD has a higher expected CRPS than QRF (average CRPSS of -1.37%), but they have improved predictive performance (in terms of CRPS) over a non-negligible part of the region of interest. When censoring grid points located over the sea and at the border, the average CRPSS w.r.t. QRF is 0.26%, showing that U-Net+CSGD has a predictive performance comparable to QRF over land in terms of CRPS.

As mentioned in Section 1.1.3, rank histograms are a useful diagnostic tool for probabilistic calibration of forecasts. In particular, the flatness of the rank histogram characterizes calibrated forecasts. Flatness and other informative shapes can be statistically tested for by, as referred to here, Jolliffe-Primo-Zamo (JPZ) tests (Jolliffe and Primo, 2008; Zamo, 2016). Figure 1.3 shows the rank histograms over the whole grid and the JPZ tests for flatness. As is often the case, the raw ensemble is biased and underdispersed, which is visible by the triangular shape of the rank histograms and the fact that the lowest and highest ranks are over-represented. Its JPZ test confirms that the raw ensemble forecast is not calibrated (only 6% of grid points do not reject the flatness of the rank histogram). QRF shows a very high calibration with JPZ tests not rejecting flatness at 93% of grid points. U-Net+CSGD methods present a lower calibration level compared to QRF, but it is still significantly calibrated. The JPZ tests do not reject the flatness hypothesis at 77% of the grid points. The grid points at which U-Net+CSGD forecasts are not calibrated (i.e., JPZ rejecting the flatness hypothesis) are associated with high climatologies.

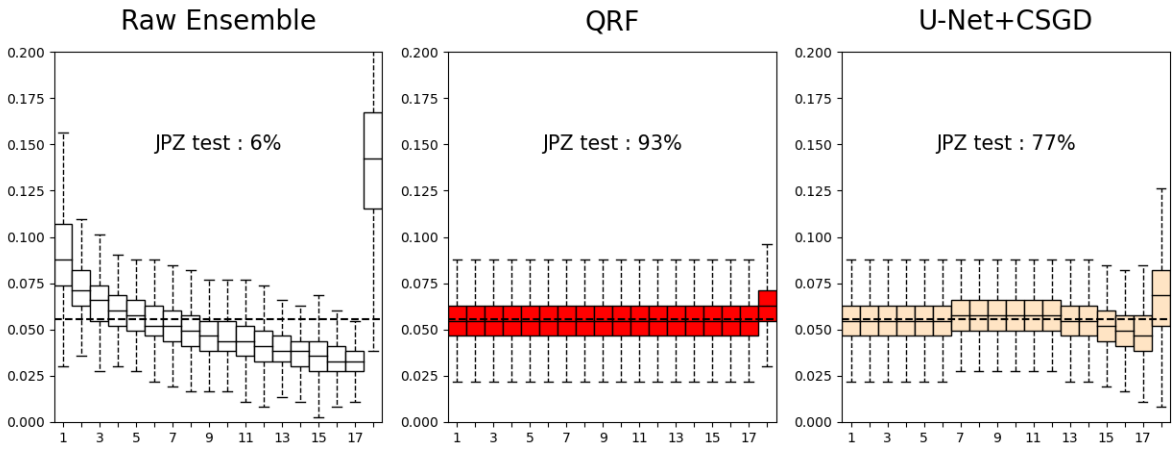


Figure 1.3: Rank histogram for the raw ensemble, QRF and U-Net+CSGD.

To focus on forecasts' predictive performance regarding extreme events, we are interested in predicting binary events in the form of the exceedance of a high threshold t (see Fig. 1.4). We use ROC (Receiver Operating Characteristic) curves to evaluate the discriminant power of forecasts in terms of binary decisions. In particular, ROC curves can inform on the risk of missing an extreme event. A good forecast should maximize the rate of events detected while minimizing false alarms. For high thresholds $t = 10$ mm and $t = 20$ mm (corresponding to the quantile of level 0.995 and 0.999, respectively, of the climatology over the region of inter-

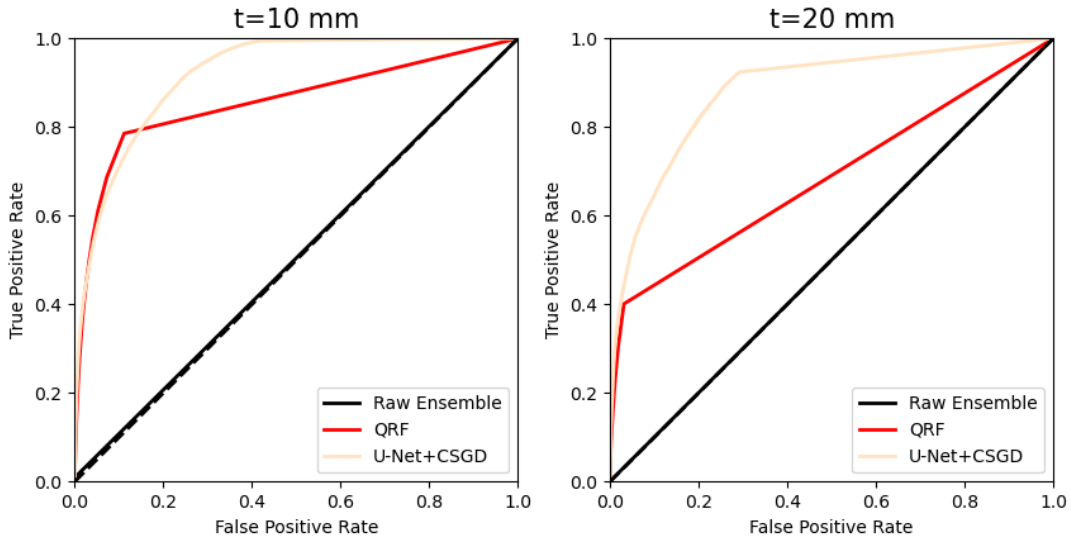


Figure 1.4: Receiver operating characteristic (ROC) curves of binary events corresponding to the exceedance of a threshold $t = 10$ mm and $t = 20$ mm.

est), the ROC curves of the different postprocessing methods have a clear ranking. For both thresholds, the performance of the raw ensemble is close to the random guess (dashed line). For $t = 10$ mm, both QRF and U-Net+CSGD are able to maintain good predictive power. However, U-Net+CSGD has a better performance than QRF. For $t = 20$ mm, the gap in performance between U-Net+CSGD and QRF continues to grow as the predictive performance of QRF deteriorates.

We propose U-Net-based methods that can simultaneously postprocess marginal distributions at each grid point using information from nearby grid points. It circumvents the storage memory voracity and the inability to extrapolate of QRF. DRU outperforms the raw ensemble for all metrics used. Moreover, DRUs have predictive performance comparable to QRF-based methods in terms of CRPS. DRUs are (probabilistically) calibrated over a large part of the domain studied except for areas associated with high climatological precipitation. Regarding predictive power for heavy precipitation, U-Net+CSGD outperforms QRF-based methods.

Chapter 3 : Summary of contributions

- Distributional regression U-Net (DRU) is a global model that predicts marginal parametric distributions and circumvents some known limitations of QRF. It provides a natural extension of DRN to grid-based data.
- We review methods using U-Net architectures in statistical postprocessing (Table 3.4).
- In terms of CRPS, the predictive performance of DRU is comparable to state-of-the-art methods (Figure 3.5 and Table 3.5).
- DRUs provide (probabilistically) calibrated forecasts at most grid points. However, it fails in areas of high climatological precipitation (Figures 3.8 and 3.9).
- U-Net+CSGD outperforms other methods in terms of exceedance of high-precipitation thresholds (Figure 3.10).

1.4 Proper scoring rules for multivariate probabilistic forecasts based on aggregation and transformation

The previous section (and the associated chapter) does not postprocess the dependence structure of the quantity of interest, assuming it can be retrieved from the raw ensemble (e.g., using ECC; Schefzik et al. 2013) or from the climatology (e.g., using ScS; Clark et al. 2004) or that it can be processed separately as mentioned in Section 1.1. Nonetheless, it is a crucial aspect of forecasts since it can influence the impact of an event. Spatial probabilistic forecasts require appropriate verification methods.

Scoring rules are a tool of choice to both quantify how good a forecast is and compare competing forecasts. Recall that propriety allows to assess both calibration and sharpness simultaneously (Winkler, 1977; Winkler et al., 1996); thus, it encourages forecasters to follow their true beliefs and prevents hedging. However, it is a necessary property of good scoring rules, but it does not guarantee that a scoring rule provides an informative characterization of predictive performance. In particular, propriety does not ensure that forecasts minimizing the expected score are of interest to the task at hand. Even strict propriety does not ensure that forecasts in the vicinity of the minimum expected score are close to the ideal forecast in an interesting manner. In univariate and multivariate settings, numerous studies have proven that no scoring rule has it all, and thus, different scoring rules should be used to get a better understanding of the predictive performance of forecasts (see, e.g., Scheuerer and Hamill 2015b; Taillardat 2021; Bjerregård et al. 2021).

This may explain the development of *spatial verification tools* (Gilleland et al., 2009; Dorninger et al., 2018) which are physics-based verification methods for spatial forecasts. They rely on robustness to the double-penalty effect (Ebert, 2008) and the interpretability of both single values and the ranking of forecasts. However, the vast majority of methods are not proper. In the context of proper scoring rules, interpretability can arise from being induced by a consistent scoring function for a functional (e.g., the squared error is induced by a scoring function consistent for the mean; Gneiting 2011), knowing what aspects of the forecast the scoring rule discriminates (e.g., the Dawid-Sebastiani score only discriminates forecasts through their mean and variance; Dawid and Sebastiani 1999) or knowing the limitations of a certain proper scoring rule (e.g., the variogram score is incapable of discriminating two forecasts that only differ by a constant bias; Scheuerer and Hamill 2015b). In this context, interpretable proper scoring rules become verification methods of choice as the ranking of forecasts they produce can be more informative than the ranking of a more complex but less interpretable scoring rule.

Scheuerer and Hamill (2015b) proposed the variogram score to target the verification of the dependence structure. The variogram score of order p ($p > 0$) is defined as

$$\mathrm{VS}_p(F, \boldsymbol{y}) = \sum_{i,j=1}^{d} w_{ij} \left(\mathbb{E}_F \left[|X_i - X_j|^p \right] - |y_i - y_j|^p \right)^2 ,$$

where X_i is the i-th component of the random vector $\boldsymbol{X} \in \mathbb{R}^d$ following F, the w_{ij} are non-negative weights and $\boldsymbol{y} \in \mathbb{R}^d$ is an observation. The construction of the variogram score relies on two main principles. First, the variogram score is the weighted sum of scoring rules acting on the distribution of $\boldsymbol{X}_{i,j} = (X_i, X_j)$ and on paired components of the observations $y_{i,j}$. This *aggregation* principle allows the combination of proper scoring rules and summarizes them into a proper scoring rule acting on the whole distribution F and observations $\boldsymbol{y}$. Second, the scoring rules composing the weighted sum can be seen as a standard proper scoring rule applied to transformations of both forecasts and observations. Let us denote $\gamma_{i,j} : \boldsymbol{x} \mapsto |x_i - x_j|^p$ the transformation related to the variogram of order p, then the variogram score can be rewritten

as

$$\mathrm{VS}_p(F, \boldsymbol{y}) = \sum_{i,j=1}^{d} w_{ij} \mathrm{SE}(\gamma_{i,j}(F), \gamma_{i,j}(\boldsymbol{y})),$$

where $\mathrm{SE}(F, y) = (\mathbb{E}_F[X] - y)^2$ is the univariate squared error (SE) and $\gamma_{i,j}(F)$ is the distribution of $\gamma_{i,j}(\boldsymbol{X})$ for $\boldsymbol{X} \sim F$. This second principle is the *transformation* principle, allowing to build transformation-based proper scoring rules that can benefit from interpretability arising from a transformation (here, the variogram transformation $\gamma_{i,j}$) and the simplicity and interoperability of the proper scoring rule they rely on (here, the SE).

These two principles have been disseminated across the literature for the past decades. More explicitly, Dawid and Musio (2014) proposes the notion of *composite score* which is a particular case of the combination of both principles. Heinrich-Mertsching et al. (2024) introduces the transformation principle and applies it in the context of point processes. We formalize general forms of the aggregation and transformation principles and their combination leads to Corollary 1.1.

Corollary 1.1. *Let $\mathcal{T} = \{T_i\}_{1 \leq i \leq m}$ be a set of transformations from $\mathbb{R}^d$ to $\mathbb{R}^k$. Let $\mathcal{S}_{\mathcal{T}} = \{\mathrm{S}_{T_i}\}_{1 \leq i \leq m}$ be a set of proper scoring rules where S is proper relative to $T_i(\mathcal{F})$, for all $1 \leq i \leq m$. Let $\boldsymbol{w} = \{w_i\}_{1 \leq i \leq m}$ be nonnegative weights. Then, the scoring rule*

$$\mathrm{S}_{\mathcal{S}_{\mathcal{T}}, \boldsymbol{w}}(F, \boldsymbol{y}) = \sum_{i=1}^{m} w_i \mathrm{S}_{T_i}(F, \boldsymbol{y})$$

is proper relative to $\mathcal{F}$.

To gain interpretability, it is natural to have dimension-reducing transformations (i.e., $k < d$) as it leads to transformations simplifying the multivariate quantities. Particularly, it is generally preferred to choose $k = 1$ to make the quantity easier to interpret and focus on specific information contained in the forecast or the observation. Additionally, we show that all kernel scoring rules can be expressed as the aggregation of SE applied to a sequence of transformations.

Aggregation-and-transformation-based scoring rules can leverage the interpretability of both transformations and standard scoring rules. For example, if interest is on the predictive performance of forecasts in terms of their prediction of the exceedance of a threshold t, the Brier score (Brier, 1950) should be used in a univariate setup. The Brier score is expressed as

$$\mathrm{BS}_t(F, y) = ((1 - F(t)) - \mathbb{1}_{y > t})^2 = (F(t) - \mathbb{1}_{y \leq t})^2,$$

where $1 - F(t)$ is the predicted probability that the threshold t is exceeded. The expectation of the Brier score is minimum for all forecasts F such that the probability of exceedance of threshold t is correctly predicted. In a spatial verification context, the exceedance of the threshold at each location can be summarized by the aggregated Brier score

$$\frac{1}{d} \sum_{i=1}^{d} \mathrm{BS}(F_i, y_i),$$

where F_i is the marginal distribution of F at the location i and y_i is the value of $\boldsymbol{Y}$ at the location i. In that case, the transformations are projections onto each location (i.e., the 1-dimensional marginals) and the aggregation uses uniform weights since no assumption is made on the locations.

When considering the spatial dependence structure, a quantity of interest can be the exceedance of a threshold t at neighboring locations. In the case of precipitation, the neighborhoods might be defined as catchment areas of rivers. The fraction of threshold exceedance (FTE)

is the summary statistic associated with the simultaneous exceedance of a certain threshold and it is defined as

$$\mathrm{FTE}_{P,t}(\boldsymbol{X}) = \frac{1}{|P|}\sum_{i\in P}\mathbb{1}_{\{X_i\geq t\}},$$

where P is a patch (or neighborhood) of interest and $|P|$ its dimension. Using the aggregation and transformation principles, the aggregated SE of FTE is defined as

$$\sum_{P\in\mathcal{P}} w_P \mathrm{SE}\big(\mathrm{FTE}_{P,t}(F), \mathrm{FTE}_{P,t}(\boldsymbol{y})\big) = \sum_{P\in\mathcal{P}} w_P \big(\mathbb{E}_F[\mathrm{FTE}_{P,t}(X)] - \mathrm{FTE}_{P,t}(\boldsymbol{y})\big)^2$$

where $\mathcal{P}$ is an ensemble of patches, w_P is the weight associated with a patch $P \in \mathcal{P}$. This scoring rule is proper and focuses on the prediction of the exceedance of a threshold t via the fraction of locations over a patch P exceeding said threshold. The resemblance with the Brier score is clear and the aggregated SE of FTE becomes the aggregated BS when patches containing a single location are considered.

Numerous other examples of transformations (and the scoring rules they result in) are presented, discussed, and linked to the literature in Chapter 4. Multiple numerical experiments are developed to showcase the importance of interpretability in a practical setting and, more particularly, how aggregation-and-transformation-based scoring rules can be used in spatial forecast verification. In particular, we show how usual scoring rules can be adapted to avoid the double-penalty effect.

Chapter 4 : Summary of contributions

- We provide a comprehensive review of both univariate and multivariate scoring rules in light of interpretability (Section 4.2).
- We formalize a framework (disseminated in the literature) based on the aggregation and transformation principles to construct interpretable proper scoring rules (Section 4.3).
- Kernel scores can be expressed as an aggregation of squared error applied to a sequence of transformations (Proposition 4.3).
- List examples of aggregation-and-transformation-based scoring rules from both the literature and original suggestions (Section 4.4).
- Numerical experiments have been conducted to illustrate the benefit of interpretable proper scoring rules in various contexts (Section 4.5).
- In particular, concrete solutions are given to help bridge the gap with spatial verification tools.

1.5 Related works

Appendix A and Appendix B reproduce two related works conducted during the thesis (Dombry et al., 2024; Al Masry et al., 2023). These works are related to distributional regression but not directly connected to statistical postprocessing. Hereafter, we briefly motivate these works, explicit their relation to the works presented in the previous sections and summarize their main contributions.

1.5.1 Stone's theorem for distributional regression in Wasserstein distance

As mentioned above, Chapter 2 adapts results from point regression to distributional regression. Instead of targeting optimal convergence for a given class of distribution, Appendix A focuses on universal consistency in distributional regression (i.e., convergence results holding for any distribution but without guarantee on the rate of convergence).

Recall the general regression framework introduced in Section 1.2. We observe a sample $D_n = \{(X_i, Y_i), 1 \leq i \leq n\}$, of independent copies of $(X, Y) \in \mathbb{R}^d \times \mathbb{R}^m$ with distribution P. Based on this sample and assuming Y integrable, in point regression, the goal is to estimate the regression function

$$r(x) = \mathbb{E}[Y|X = x], \quad x \in \mathbb{R}^d.$$

Local average estimators take the form

$$\hat{r}_n(x) = \sum_{i=1}^{n} W_{ni}(x) Y_i$$

with $W_{n1}(x), \ldots, W_{nn}(x)$ the *local weights* at x. The local weights are assumed to be measurable functions of x and $X_1, \ldots, X_n$ but not to depend on $Y_1, \ldots, Y_n$. Consider the case of probability weights satisfying

$$W_{ni}(x) \geq 0, \ 1 \leq i \leq n, \quad \text{and} \quad \sum_{i=1}^{n} W_{ni}(x) = 1. \tag{1.7}$$

Stone's Theorem states the universal consistency of the regression estimate in L^p-norm.

Theorem 1.2 (Stone 1977)**.** *Assume the probability weights* (1.7) *satisfy the following three conditions:*

i) there is $C > 0$ *such that* $\mathbb{E}\big[\sum_{i=1}^{n} W_{ni}(X) g(X_i)\big] \leq C\mathbb{E}[g(X)]$ *for all* $n \geq 1$ *and measurable* $g : \mathbb{R}^k \to [0, +\infty)$ *such that* $\mathbb{E}[g(X)] < \infty$*;*

ii) for all $\varepsilon > 0$, $\sum_{i=1}^{n} W_{ni}(X) \mathbb{1}_{\{\|X_i - X\| > \varepsilon\}} \to 0$ *in probability as* $n \to +\infty$*;*

iii) $\max_{1 \leq i \leq n} W_{ni}(X) \to 0$ *in probability as* $n \to +\infty$*.*

Then, for all $p \geq 1$ *and* $(X, Y) \sim P$ *such that* $\mathbb{E}[\|Y\|^p] < \infty$*,*

$$\mathbb{E}\big[\|\hat{r}_n(X) - r(X)\|^p\big] \longrightarrow 0 \quad \text{as } n \to +\infty. \tag{1.8}$$

Conversely, if Equation (1.8) *holds, then the probability weights must satisfy conditions* $i) - iii)$*.*

Examples of local average estimators include k-NN, kernel methods and some variants of random forests.

To adapt this result to distributional regression, we use a definition of convergence based on Wasserstein distance rather than relying on scoring rules, contrary to Chapter 2 :

$$\mathbb{E}\left[\mathcal{W}_p^p(\hat{F}_{n,X}, F_X^*)\right] \longrightarrow 0 \quad \text{as } n \to +\infty, \tag{1.9}$$

where $\mathcal{W}_p$ is the Wasserstein distance of order p and $\hat{F}_{n,X}$ is an estimate of the conditional distribution of Y given X, noted F_X^*. Consider the weighted empirical distribution estimator based on the training sample D_n

$$\hat{F}_{n,X} = \sum_{i=1}^{n} W_{ni}(X) \delta_{Y_i}, \tag{1.10}$$

where δ_y denotes the Dirac mass at point $y \in \mathbb{R}^m$.

Using the notion of max-sliced Wasserstein distance (Bayraktar and Guo, 2021), this work extends Stone's theorem to distributional regression in Wasserstein distance of order $p \geq 1$. More precisely, there is an equivalence between the fact that the weights of the weighted empirical distribution estimator (1.10) satisfies conditions $i) - iii)$ and the convergence of (1.9). Moreover, for $p = 1$, the optimal minimax rates of convergence on specific classes of distributions are obtained. Applications of Stone's theorem in distributional regression are illustrated using the estimation of conditional tail expectation and probability-weighted moments, among others.

Appendix A : Summary of contributions

- We adapt Stone's theorem to (multivariate) distributional regression: universal consistency in terms of Wasserstein distance of order $p \geq 1$ in a multivariate setting (Theorem A.2).
- We determine optimal minimax rates of convergence for $p = 1$ and $m = 1$ (Theorem A.3).
- We illustrate applications of Stone's theorem to summary statistics in distributional regression (Section A.3.3).

1.5.2 A new methodology to predict the oncotype scores based on clinico-pathological data with similar tumor profiles

Appendix B uses a distributional regression technique to assist clinicians in their decision-making regarding the prediction of breast cancer recurrence risk and potential treatments.

The Oncotype DX (ODX) test is a commercially available molecular test for breast cancer assay that provides prognostic and predictive breast cancer recurrence information for hormone-positive, HER2-negative patients. The ODX test provides a recurrence score (ODX score) between 0 and 100. Higher values of the ODX score correspond to a higher risk of recurrence. Several retrospective and prospective studies have validated this test and its clinical utility (see, e.g., Paik et al. 2004, 2006; Albain et al. 2010). The most common interpretations of the ODX score are through cutoffs defining two or three classes of risk: for example, low risk < 11, intermediate risk 11-25 and high risk > 25 (Sparano et al., 2018). Despite its clinical utility, the ODX test is expensive and the ODX score lacks explainability. In order to bypass these limitations, studies have tried to use clinico-pathological characteristics to predict the ODX score via its direct value or via a classification in terms of risk levels. Numerous statistical learning methods have been studied, such as multiple linear regression (Klein et al., 2013; Hou et al., 2017), random forests (Kim et al., 2019; Pawloski et al., 2021) and neural networks (Kim et al., 2019; Baltres et al., 2020).

In order to have complete knowledge of the uncertainty, we proposed to predict the full distribution of the ODX conditionally on clinico-pathological characteristics: the distributional regression forest (DRF; Meinshausen 2006; Athey et al. 2019). As DRF provides a probabilistic prediction, its output can take the form of a discrete probability density function. Moreover, it can be summarized by more understandable quantities for practitioners, such as its mean and the standard deviation or the probabilities of classes of interest, leveraging the practitioner's familiarity with interpretations of the ODX score. In particular, patients with similar profiles (in terms of the weights of the forest) can be used to inform practitioners of related patients present in the cohort and detect uninformative predictions linked to a lack of representativity. In the meantime, DRF has a performance comparable to the previous methods proposed in the

literature.

Appendix B : Summary of contributions

- We use a distributional regression technique (DRF) to predict breast cancer recurrence risk and provide information useful for decision-making.
- DRFs provide a probabilistic prediction that can be summarized in understandable quantities for practitioners (e.g., a neighborhood of close patients on the cohort) (Figure B.2).
- In terms of low-risk and high-risk classification, DRFs achieve a predictive performance comparable to state-of-the-art methods (Table B.3).

1.6 Publications list

This thesis includes the following articles written during the study period :

- R. Pic, C. Dombry, P. Naveau, M. Taillardat (2023). "Distributional regression and its evaluation with the CRPS: Bounds and convergence of the minimax risk." *International Journal of Forecasting*, 39(4), pp. 1564–1572. doi:10.1016/j.ijforecast.2022.11.001
- R. Pic, C. Dombry, P. Naveau, M. Taillardat (2024). "Distributional regression U-Nets for the postprocessing of precipitation ensemble forecasts." Submitted to *Artificial Intelligence for the Earth Systems*. arXiv, HAL
- R. Pic, C. Dombry, P. Naveau, M. Taillardat (2024). "Proper scoring rules for multivariate probabilistic forecasts based on aggregation and transformation." Submitted to *Advances in Statistical Climatology, Meteorology and Oceanography*. arXiv, HAL
- C. Dombry, T. Modeste, R. Pic (2024). "Stone's theorem for distributional regression in Wasserstein distance." *Journal of Nonparametric Statistics*, pp. 1-23. doi:10.1080/10485252.2024.2393172
- Z. Al Masry, R. Pic, C. Dombry, C. Devalland (2024). "A new methodology to predict the oncotype scores based on clinico-pathological data with similar tumor profiles." *Breast Cancer Research and Treatment*, 203, pp. 587–598. doi:10.1007/s10549-023-07141-5

Chapter 2

Distributional regression and its evaluation with the CRPS: bounds and convergence of the minimax risk

This chapter reproduces an article published in *International Journal of Forecasting*, and written by Romain Pic[1], Clément Dombry[1], Philippe Naveau[2] and Maxime Taillardat[3].

Abstract The theoretical advances on the properties of scoring rules over the past decades have broadened the use of scoring rules in probabilistic forecasting. In meteorological forecasting, statistical postprocessing techniques are essential to improve the forecasts made by deterministic physical models. Numerous state-of-the-art statistical postprocessing techniques are based on distributional regression evaluated with the Continuous Ranked Probability Score (CRPS). However, theoretical properties of such evaluation with the CRPS have solely considered the unconditional framework (i.e. without covariates) and infinite sample sizes. We extend these results and study the rate of convergence in terms of CRPS of distributional regression methods. We find the optimal minimax rate of convergence for a given class of distributions and show that the k-nearest neighbor method and the kernel method reach this optimal minimax rate.

Contents

[1]Université de Franche Comté, CNRS, LmB (UMR 6623), F-25000 Besançon, France
[2]Laboratoire des Sciences du Climat et de l'Environnement, UMR 8212, CEA-CNRS-UVSQ, EstimR, IPSL & U Paris-Saclay, Gif-sur-Yvette, France
[3]CNRM, Université de Toulouse, Météo-France, CNRS, Toulouse, France

2.1 Introduction

In meteorology, ensemble forecasts are based on a given number of deterministic models whose parameters vary slightly in order to consider observation errors and incomplete physical representation of the atmosphere. This leads to an ensemble of different forecasts that overall also assess the uncertainty of the forecast. Ensemble forecasts suffer from bias and underdispersion (Hamill and Colucci, 1997; Baran and Lerch, 2018) and need to be statistically postprocessed in order to be improved. Different postprocessing methods have been proposed, such as Ensemble Model Output Statistics (Gneiting et al., 2005), Quantile Regression Forests (Taillardat et al., 2016) or Neural Networks (Schulz and Lerch, 2022b). These references, among others, also discuss the stakes of weather forecast statistical postprocessing.

Postprocessing methods rely on distributional regression (Gneiting and Katzfuss, 2014) where the aim is to predict the conditional distribution of the quantity of interest (e.g. temperatures, wind speed, or precipitation) given a set of covariates (e.g. raw outputs of a physical ensemble model). Algorithms are often based on the minimization of a proper scoring rule that compares actual observations with the predictive distribution. Scoring rules can be seen as an equivalent of loss functions in classical regression. A detailed review of scoring rules is given by Gneiting and Raftery (2007). The Continuous Ranked Probability Score (CRPS; Matheson and Winkler, 1976), defined in Equation (2.2), is one of the most popular scores in meteorological forecasting. The CRPS is also minimized to infer parameters of statistical models used in postprocessing (e.g. Gneiting et al., 2005; Naveau et al., 2016; Rasp and Lerch, 2018; Taillardat et al., 2019). Recently, under monotonicity assumptions, the isotonic distributional regression Henzi et al. (2021) was shown to minimize the in-sample CRPS and to satisfy consistency in the sense of Kolmogorov distance.

To the best of our knowledge, most convergence statements in distributional regression (e.g. Thorey et al., 2017 and Mösching and Dümbgen, 2020) are not only derived within an unconditional framework, i.e. without taking into account the covariates, but also these limiting results assume arbitrarily large sample sizes. In this work, our goal is to bypass these two limitations.

This paper is organized as follows. Section 2.2 introduces preliminary notions that are needed to state our main results in Section 2.3. Section 2.2.1 introduces our framework and notation for distributional regression. Section 2.2.2 provides the theoretical background on distributional regression and its evaluation using the CRPS and Section 2.2.3 provides some elements on minimax risk theory. Section 2.2.4 briefly introduces the two models that are studied in this article: the k-nearest neighbor and kernel estimators. The main result on minimax rate of convergence for distributional regression is stated in Section 2.3.1 where suitable classes of distributions $\mathcal{D}^{(h,C,M)}$ are defined. In Section 2.3.2, we study the k-NN estimators and derive a non-asymptotic upper bound for the excess risk of the CRPS uniformly on the class $\mathcal{D}^{(h,C,M)}$. Section 2.3.3 provides similar results for the kernel method. In Section 2.3.4, we find a lower minimax rate of convergence by reducing the problem to standard point regression solved by Györfi et al. (2002). We can deduce that the k-NN method for the distributional regression

reaches the optimal rate of convergence in dimension $d \geq 2$, while the kernel method reaches the optimal rate of convergence in any dimension. All the proofs are postponed to and detailed in the Appendix. A short conclusion and discussion is provided in Section 2.4.

2.2 Preliminaries

2.2.1 Distributional regression framework

In this article, we consider the regression framework $(X, Y) \in \mathbb{R}^d \times \mathbb{R}$ with distribution P. The goal of distributional regression is to estimate the conditional distribution of Y given $X = x$, noted

$$F_x^*(y) \colon = P(Y \leq y | X = x), \quad x \in \mathbb{R}^d.$$

In forecast assessment, we make the distinction between the construction of the estimator relying on the training sample $D_n = \{(X_i, Y_i), 1 \leq i \leq n\}$ and its evaluation with respect to new data (X, Y). Given the training sample D_n, the forecaster constructs a predictor $\hat{F}_n : x \mapsto \hat{F}_{n,x}$ that estimates the conditional distribution F_x^*. In this context, it is crucial to assess if $\hat{F}_{n,x}$ is close to F_x^* over the entire range of possible values of $X = x$. To this aim, we consider

$$\mathbb{E}_{X \sim P_X, D_n \sim P^n} \left[\int_{\mathbb{R}} |\hat{F}_{n,X}(z) - F_X^*(z)|^2 \mathrm{d}z \right] \tag{2.1}$$

where P_X denotes the marginal distribution of X, $\mathbb{E}_{X \sim P_X, D_n \sim P^n}$ denotes the expectation with respect to X and D_n following P_X and P^n respectively. The squared L^2-norm within the expectation is usually referred to as the squared second-order *Cramér's distance*. We focus on this specific distance because it corresponds to the excess risk associated with the CRPS, also called *divergence* of the CRPS, as explained in the next section.

2.2.2 CRPS and evaluation of distributional regression

The Continuous Ranked Probability Score (CRPS; Matheson and Winkler, 1976) compares a predictive distribution F and a real-valued observation y by computing the following integral

$$\mathrm{CRPS}(F, y) = \int_{\mathbb{R}} (F(z) - \mathbb{1}_{y \leq z})^2 \mathrm{d}z. \tag{2.2}$$

The expected CRPS of a predictive distribution F when the observations Y are distributed according to G is defined as

$$\overline{\mathrm{CRPS}}(F, G) = \int_{\mathbb{R}} \mathrm{CRPS}(F, y) G(\mathrm{d}y), \quad F, G \in \mathcal{M}(\mathbb{R}), \tag{2.3}$$

where $\mathcal{M}(\mathbb{R})$ denotes the set of all distribution functions on $\mathbb{R}$. This quantity is finite when both F and G have a finite first moment. Then, the difference between the expected CRPS of the forecast F and the expected CRPS of the ideal forecast G can be written as

$$\overline{\mathrm{CRPS}}(F, G) - \overline{\mathrm{CRPS}}(G, G) = \int_{\mathbb{R}} |F(z) - G(z)|^2 \mathrm{d}z \geq 0. \tag{2.4}$$

This implies that the only optimal prediction, in the sense that it minimizes the expected CRPS, is the true distribution G. A score with this property is said to be *strictly proper*. This property is essential for distributional regression as it justifies the minimization of the expected score in order to construct or evaluate a prediction.

In distributional regression, the quality of a predictor $\hat{F} : x \mapsto \hat{F}_x$ is assessed by its risk

$$\begin{aligned} R_P(\hat{F}) &= \mathbb{E}_{(X,Y)\sim P}\left[\mathrm{CRPS}(\hat{F}_X, Y)\right] \\ &= \mathbb{E}_{X\sim P_X}\left[\overline{\mathrm{CRPS}}(\hat{F}_X, F^*_X)\right]. \end{aligned}$$

This quantity is important as many distributional regression methods try to minimize it in order to improve predictions. When Y is integrable, Equation (2.4) implies

$$\begin{aligned} R_P(\hat{F}) - R_P(F^*) &= \mathbb{E}_{(X,Y)\sim P}\left[\mathrm{CRPS}(\hat{F}_X, Y) - \mathrm{CRPS}(F^*_X, Y)\right] \\ &= \mathbb{E}_{X\sim P_X}\left[\int_{\mathbb{R}} \left|\hat{F}_X(z) - F^*_X(z)\right|^2 \mathrm{d}z\right] \geq 0. \end{aligned} \tag{2.5}$$

We recall that the Bayes risk is the minimal theoretical risk over all possible predictors and that a Bayes predictor is a predictor achieving the Bayes risk. Thus, Equation (2.5) implies that $R_P(F^*)$ is the Bayes risk and that F^* is a Bayes predictor if and only if $\hat{F}_x = F^*_x$ P_X-a.e. An introduction to the notions of theoretical risk, Bayes risk and excess risk can be found in Section 2.4 of Hastie et al. (2009).

Finally, we consider the case of a predictor $\hat{F}_n$ built on a training sample $D_n = \{(X_i, Y_i), 1 \leq i \leq n\}$, as presented in Section 2.2.1, to estimate the conditional distribution of Y given X. Then, (X, Y) denotes a new independent observation used to evaluate the performances of $\hat{F}_n$. The predictor has the expected CRPS

$$\mathbb{E}_{D_n\sim P^n}[R_P(\hat{F}_n)] = \mathbb{E}_{D_n\sim P^n,(X,Y)\sim P}[\mathrm{CRPS}(\hat{F}_{n,X}, Y)],$$

with expectation taken both with respect to the training sample D_n and test observation (X, Y). Once again, when Y is integrable, the theoretical risk has a unique minimum given by $R_P(F^*)$. The *excess risk* becomes

$$\begin{aligned} &\mathbb{E}_{D_n\sim P^n}\left[R_P(\hat{F}_n)\right] - R_P(F^*) \\ &= \mathbb{E}_{D_n\sim P^n, X\sim P_X}\left[\int_{\mathbb{R}} \left|\hat{F}_{n,X}(z) - F^*_X(z)\right|^2 \mathrm{d}z\right] \geq 0. \end{aligned} \tag{2.6}$$

This justifies the choice of the squared Cramér's distance in Equation (2.1).

For large sample sizes, one expects that the predictor correctly estimates the conditional distribution and that the excess risk (2.6) tends to zero. A genuine question is to investigate the rate of convergence of the excess risk to zero as the sample size $n \to \infty$. The risk depends on the distribution of observations and we want the model to perform well on large classes of distributions. Hence, we consider the standard minimax approach, as described in the next section.

2.2.3 Optimal minimax rates of convergence

In order to study the rate of convergence, as $n \to \infty$, of the excess risk (2.6) to zero, we introduce the notion of *optimal minimax rate of convergence*. The minimax risk corresponds to the best achievable risk in the worst-case scenario (whence the name minimax). More precisely, given a class of distributions $\mathcal{D}$, the optimal minimax rate of convergence quantifies the minimal error that an estimator $\hat{F}_n$ can achieve uniformly on a given class of distributions $\mathcal{D}$, when the size of the training set D_n gets large.

Stone (1982) provided minimax rates of convergence within a point regression framework and the minimax theory for nonparametric regression is well-developed, see e.g. Györfi et al. (2002) or Tsybakov (2009). To the extent of our knowledge, this paper states the first results for distributional regression.

The formal definition of minimax rate of convergence for distributional regression is as follows.

Definition 2.1. *A sequence of positive numbers (a_n) is called an optimal minimax rate of convergence on the class $\mathcal{D}$ if*

$$\liminf_{n\to\infty} \inf_{\hat{F}_n} \sup_{P\in\mathcal{D}} \frac{\mathbb{E}_{D_n\sim P^n}[R_P(\hat{F}_n)] - R_P(F^*)}{a_n} > 0 \tag{2.7}$$

and

$$\limsup_{n\to\infty} \inf_{\hat{F}_n} \sup_{P\in\mathcal{D}} \frac{\mathbb{E}_{D_n\sim P^n}[R_P(\hat{F}_n)] - R_P(F^*)}{a_n} < \infty, \tag{2.8}$$

where the infimum is taken over all distributional regression models $\hat{F}_n$ trained on D_n. If the sequence (a_n) satisfies only the lower bound (2.7), it is called a lower minimax rate of convergence.

2.2.4 k-NN and kernel predictors in distributional regression

Many predictors $\hat{F}_n$ can be studied and possibly achieve the optimal minimax rate of convergence. In this paper, we focus on two simple cases: k-nearest neighbor and kernel estimators.

The k-nearest neighbor (k-NN) method is well-known in the classical framework of regression and classification (see, e.g. Biau and Devroye, 2015). In distributional regression, the k-NN method can be suitably adapted to estimate the conditional distribution F_x^* and the estimator is written as

$$\hat{F}_{n,x}(z) = \frac{1}{k_n}\sum_{i=1}^{k_n} \mathbb{1}_{Y_{i:n}(x)\leq z}, \tag{2.9}$$

where $1 \leq k_n \leq n$ and $Y_{i:n}(x)$ denotes the observation at the i-th nearest neighbor of x. As usual, possible ties are broken at random to define nearest neighbors. Note that, in weather forecast statistical postprocessing, the k-NN method corresponds to a type of analog ensemble method (see Delle Monache et al., 2013).

The kernel estimate in distributional regression (see, e.g., Chapter 5 of Györfi et al., 2002) can be expressed as

$$\hat{F}_{n,x}(z) = \frac{\sum_{i=1}^n K(\frac{x-X_i}{h_n})\mathbb{1}_{Y_i\leq z}}{\sum_{i=1}^n K(\frac{x-X_i}{h_n})}, \tag{2.10}$$

where the function $K : \mathbb{R}^d \to [0,\infty)$ is a density function, called kernel, and $h_n > 0$ is the so-called bandwidth, that depends on the sample size n. If the denominator in (2.10) vanishes, we use the convention $\hat{F}_{n,x}(z) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{Y_i\leq z}$.

Minimax rates of convergence of the k-NN and kernel models in point regression are well-studied and it is known that, for suitable choices of the number of neighbors k_n and bandwidth h_n respectively, the methods are minimax rate optimal on classes of distributions with Lipschitz or more generally Hölder continuous regression functions (see e.g. Theorem 14.5 in Biau and Devroye, 2015 and Theorem 5.2 in Györfi et al., 2002). For suitable classes of distributions defined hereafter, we are able to extend these results to distributional regression. Moreover, we obtain non-asymptotic bounds for the minimax rate of convergence for both the k-NN and kernel models (see Sections 2.3.2 and 2.3.3).

2.3 Main results

2.3.1 Optimal minimax rate of convergence

We consider the following classes of distributions.

Definition 2.2. *For $h \in (0,1]$, $C > 0$ and $M > 0$, let $\mathcal{D}^{(h,C,M)}$ be the class of distributions P such that $F_x^*(y) = P(Y \leq y|X = x)$ satisfies:*

i) $X \in [0,1]^d$ P_X*-a.s.;*

ii) *For all* $x \in [0,1]^d$, $\int_{\mathbb{R}} F_x^*(z)(1-F_x^*(z))\mathrm{d}z \leq M$*;*

iii) $\|F_{x'}^* - F_x^*\|_{L^2} \leq C\|x'-x\|^h$ *for all* $x, x' \in [0,1]^d$.

Conditions $i) - iii)$ in Definition 2.2 are very similar to the conditions considered in the point regression framework, see Theorem 5.2 in Györfi et al. (2002). In condition $i)$, $[0,1]^d$ could be replaced by any compact set of $\mathbb{R}^d$. Condition $ii)$ requires that $\overline{\mathrm{CRPS}}(F_x^*, F_x^*)$ remains uniformly bounded by M, which is a condition on the dispersion of the distribution F_X^* since it implies that the absolute mean error (MAE) remains uniformly bounded. Condition $iii)$ is a regularity statement of the conditional distribution in the space $L^2(\mathbb{R})$. As an illustration, the different conditions are expressed for the Generalized Pareto distribution model in Section 2.3.5 below.

Our main result is the following optimal minimax rate of convergence.

Theorem 2.1. *The sequence* $a_n = n^{-\frac{2h}{2h+d}}$ *is the optimal minimax rate of convergence on the class* $\mathcal{D}^{(h,C,M)}$.

It should be stressed that the rate of convergence $n^{-\frac{2h}{2h+d}}$ is the same as in point regression with square error, see Theorems 3.2 and 5.2 in Györfi et al. (2002) for the lower bound and upper bound, respectively.

Remark 2.1. As pointed out by a referee, conditions *i)* and *iii)* together with the integrability of Y imply condition *ii)* for some $M > 0$. However, the dispersion, as measured by M, plays an important role throughout the proofs and, for this reason, we keep condition *ii)* in order to obtain bounds as tight as possible.

The proof of Theorem 2.1 is divided into three steps:

1. We provide in Section 2.3.2 an explicit and non-asymptotic upper bound for the excess risk of the k-nearest neighbor model uniformly on the class $\mathcal{D}^{(h,C,M)}$; the upper bound is then optimized with a suitable choice of $k = k_n$.

2. In Section 2.3.3, we obtain similar results for the kernel model.

3. We show in Section 2.3.4 that $a_n = n^{-\frac{2h}{2h+d}}$ is a lower minimax rate of convergence; the main argument is that it is enough to consider a binary model when both the observation Y and prediction $\hat{F}_X$ take values in $\{0, L\}$; we deduce that in this case, the CRPS coincides with the mean squared error so that we can appeal to standard results on lower minimax rate of convergence for regression.

Combining these three steps, we finally obtain Theorem 2.1 providing the optimal minimax rate of convergence of the excess risk on the class $\mathcal{D}^{(h,C,M)}$. All the proofs are postponed to the Appendix (Section 2.5).

2.3.2 Upper bound for the k-nearest neighbor model

The k-NN method for distributional regression is defined in Equation (2.9). Here we do not use only the mean of the nearest neighbor sample $(Y_{i:n}(x))_{1\leq i\leq k_n}$ but its entire empirical distribution. Interestingly, the tools developed to analyze the k-NN in point regression can be used in our distributional regression framework.

Proposition 2.1. *Assume $P \in \mathcal{D}^{(h,C,M)}$ and let $\hat{F}_n$ be the k-nearest neighbor model defined by Equation (2.9). Then,*

$$\mathbb{E}_{D_n \sim P^n}[R_P(\hat{F}_n)] - R_P(F^*) \leq \begin{cases} 8^h C^2 \left(\frac{k_n}{n}\right)^h + \frac{M}{k_n} & \text{if } d = 1, \\ {c_d}^h C^2 \left(\frac{k_n}{n}\right)^{2h/d} + \frac{M}{k_n} & \text{if } d \geq 2, \end{cases}$$

where $c_d = \frac{2^{3+\frac{2}{d}}(1+\sqrt{d})^2}{V_d^{2/d}}$ and V_d is the volume of the unit ball in $\mathbb{R}^d$.

Let us stress that the upper bound is non-asymptotic and holds for all fixed n and k_n. Optimizing the upper bound in k_n yields the following corollary.

Corollary 2.1. *Assume $P \in \mathcal{D}^{(h,C,M)}$ and consider the k-NN model (2.9).*

- *For $d = 1$, the optimal choice $k_n = \left(\frac{M}{hC^2 8^h}\right)^{\frac{1}{h+1}} n^{\frac{h}{h+1}}$ yields*

$$\mathbb{E}_{D_n \sim P^n}[R_P(\hat{F}_n)] - R_P(F^*) \leq B n^{-\frac{h}{h+1}}$$

  with constant $B = C^{\frac{2}{h+1}} M^{\frac{h}{h+1}} 8^{\frac{h}{h+1}} \left(h^{-\frac{h}{h+1}} + h^{\frac{1}{h+1}}\right)$.

- *For $d \geq 2$, the optimal choice $k_n = \left(\frac{Md}{2hC^2 c_d^h}\right)^{\frac{d}{2h+d}} n^{\frac{2h}{2h+d}}$ yields*

$$\mathbb{E}_{D_n \sim P^n}[R_P(\hat{F}_n)] - R_P(F^*) \leq B n^{-\frac{2h}{2h+d}}$$

  with constant $B = (C^2 c_d^h)^{\frac{d}{2h+d}} M^{\frac{2h}{2h+d}} \left(\left(\frac{d}{2h}\right)^{\frac{2h}{2h+d}} + \left(\frac{2h}{d}\right)^{\frac{d}{2h+d}}\right)$.

2.3.3 Upper bound for the kernel model

Kernel methods adapted to distributional regression are defined in Equation (2.10). For convenience and simplicity of notations, we develop our result for the simple uniform kernel $K(x) = \mathbb{1}_{\{\|x\| \leq 1\}}$. However, it should be stressed that all the results can be extended to boxed kernels (Györfi et al., 2002, Figure 5.7 p73) to the price of some extra multiplicative constants. For the uniform kernel, the estimator writes

$$\hat{F}_{n,x}(z) = \frac{\sum_{i=1}^n \mathbb{1}_{\{\|X_i - x\| \leq h_n\}} \mathbb{1}_{\{Y_i \leq z\}}}{\sum_{i=1}^n \mathbb{1}_{\{\|X_i - x\| \leq h_n\}}}, \tag{2.11}$$

when the denominator is non-zero and $\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{Y_i \leq z\}}$ otherwise.

Proposition 2.2. *Assume $P \in \mathcal{D}^{(h,C,M)}$ and let $\hat{F}_n$ be the kernel model defined by Equation (2.11). Then,*

$$\mathbb{E}_{D_n \sim P^n}[R_P(\hat{F}_n)] - R_P(F^*) \leq \tilde{c}_d \frac{2M + C^2 d^h + \frac{M}{n}}{n h_n^d} + C^2 h_n^{2h}$$

where $\tilde{c}_d$ only depends on d.

Once again, the upper bound is non-asymptotic and holds for all fixed n and h_n. Optimizing the upper bound in h_n yields the following corollary.

Corollary 2.2. *Assume $P \in \mathcal{D}^{(h,C,M)}$ and consider the kernel model (2.11). For any d, the optimal choice*

$$h_n = \left(\frac{\tilde{c}_d d(2M + C^2 d^h + \frac{M}{n})}{2hC^2} \right)^{\frac{1}{2h+d}} n^{-\frac{1}{2h+d}}$$

yields

$$\mathbb{E}_{D_n \sim P^n}[R_P(\hat{F}_n)] - R_P(F^*) \leq B n^{-\frac{2h}{2h+d}}$$

with

$$B = C^{\frac{2d}{2h+d}} \left(\tilde{c}_d(2M + C^2 d^h + \frac{M}{n}) \right)^{\frac{2h}{2h+d}} \left(\left(\frac{d}{2h} \right)^{-\frac{d}{2h+d}} + \left(\frac{d}{2h} \right)^{\frac{2h}{2h+d}} \right).$$

2.3.4 Lower minimax rate of convergence

We finally compare the rates of convergence obtained in Corollaries 2.1 and 2.2 with a lower minimax rate of convergence in order to see whether the optimal rate of convergence is achieved.

To prove a lower bound on a class $\mathcal{D}$, it is always possible to consider a smaller class $\mathcal{B}$. Indeed, if $\mathcal{B} \subset \mathcal{D}$, we clearly have

$$\inf_{\hat{F}_n} \sup_{P \in \mathcal{B}} \left\{ \mathbb{E}_{D_n \sim P^n}[R_P(\hat{F}_n)] - R_P(F^*) \right\} \leq \inf_{\hat{F}_n} \sup_{P \in \mathcal{D}} \left\{ \mathbb{E}_{D_n \sim P^n}[R_P(\hat{F}_n)] - R_P(F^*) \right\}$$

so that any lower minimax rate of convergence on $\mathcal{B}$ is also a lower minimax rate of convergence on $\mathcal{D}$.

To establish the lower minimax rate of convergence, we focus on the following classes of binary responses.

Definition 2.3. *Let $\mathcal{B}^{(h,C,L)}$ be the class of distributions of (X, Y) such that:*

i) $Y \in \{0, L\}$ and X is uniformly distributed on $[0,1]^d$;

ii) $\|F^*_{x'} - F^*_x\|_{L^2} \leq C\|x' - x\|^h$ for all $x, x' \in [0,1]^d$.

Since a binary outcome $Y \in \{0, L\}$ satisfies $\int_{\mathbb{R}} F^*_x(z)(1 - F^*_x(z))\mathrm{d}z \leq L/4$, condition *ii)* in Definition 2.2 holds with $M \geq L/4$. Then $\mathcal{B}^{(h,C,L)} \subset \mathcal{D}^{(h,C,M)}$ and the following lower bound established on the smaller class also holds on the larger class.

Proposition 2.3. *The sequence $a_n = n^{-\frac{2h}{2h+d}}$ is a lower minimax rate of convergence on the class $\mathcal{B}^{(h,C,L)}$. More precisely,*

$$\liminf_{n \to \infty} \inf_{\hat{F}_n} \sup_{P \in \mathcal{B}^{(h,C,L)}} \frac{\mathbb{E}_{D_n \sim P^n}[R_P(\hat{F}_n)] - R_P(F^*)}{C^{\frac{2d}{2h+d}} n^{-\frac{2h}{2h+d}}} \geq C_1 \tag{2.12}$$

for some constant $C_1 > 0$ independent of C.

Combining Corollaries 2.1 and 2.2 and Proposition 2.3, we can deduce that for $d \geq 2$, the k-NN model reaches the minimax lower rate of convergence $a_n = n^{-\frac{2h}{2h+d}}$ for the class $\mathcal{D}^{(h,C,M)}$ and that the kernel model reaches the minimax lower rate of convergence a_n in any dimension d. This shows that this lower rate of convergence is in fact the optimal rate of convergence and proves Theorem 2.1.

2.3.5 Generalized Pareto distributions

Explicit parametric formulas of the CRPS exist for most classical distribution families: e.g. Gaussian, logistic, censored logistic, Generalized Extreme Value, Generalized Pareto (see Gneiting et al., 2005; Taillardat et al., 2016; Friederichs and Thorarinsdottir, 2012). We focus here on the Generalized Pareto Distribution (GPD) family and we denote by $H_{\xi,\sigma}$ the GP distribution with shape parameter $\xi \in \mathbb{R}$ and scale parameter $\sigma > 0$. Recall that it is defined, when $\xi \neq 0$, by

$$H_{\xi,\sigma}(z) = 1 - \left(1 + \frac{\xi z}{\sigma}\right)_+^{-1/\xi}, \quad z > 0,$$

with the notation $(\cdot)_+ = \max(0, \cdot)$. When $\xi = 0$, the standard limit by continuity is used. For $\xi < 1$, the GPD has a finite first moment and the associated CRPS is given by Friederichs and Thorarinsdottir (2012)

$$\begin{aligned} &\mathrm{CRPS}\left(H_{\xi,\sigma}, y\right) \\ &= \left(y + \frac{\sigma}{\xi}\right)\left(2H_{\xi,\sigma}(y) - 1\right) - \frac{2\sigma}{\xi(\xi-1)}\left(\frac{1}{\xi - 2} + \left(1 - H_{\xi,\sigma}(y)\right)\left(1 + \xi\frac{y}{\sigma}\right)\right). \end{aligned} \tag{2.13}$$

When $Y \sim H_{\xi^*,\sigma^*}$, the expected CRPS is (Taillardat et al., 2023)

$$\begin{aligned} &\overline{\mathrm{CRPS}}\left(H_{\xi,\sigma}, H_{\xi^*,\sigma^*}\right) \\ &= \frac{\sigma^*}{1-\xi^*} + \frac{2\sigma}{1-\xi}m_0 + \frac{2\xi}{1-\xi}m_1 + 2\sigma\left(\frac{1}{1-\xi} - \frac{1}{2(2-\xi)}\right) \end{aligned} \tag{2.14}$$

with

$$m_0 = \mathbb{E}_{Y\sim H_{\xi^*,\sigma^*}}\left[\left(1 + \frac{\xi}{\sigma}Y\right)^{-1/\xi}\right], \quad m_1 = \mathbb{E}_{Y\sim H_{\xi^*,\sigma^*}}\left[Y\left(1 + \frac{\xi}{\sigma}Y\right)^{-1/\xi}\right].$$

In particular,

$$\overline{\mathrm{CRPS}}\left(H_{\xi^*,\sigma^*}, H_{\xi^*,\sigma^*}\right) = \frac{\sigma^*}{(2-\xi^*)(1-\xi^*)}.$$

We now consider the distributional regression framework and we illustrate the statement of Section 2.2.2 on Bayes risk in the case of a Generalized Pareto regression model where Y given $X = x$ follows a GPD with shape parameter $\xi^*(x)$ and scale parameter $\sigma^*(x)$. Then, it is possible to show that Bayes risk is equal to

$$R_P(F^*) = \int_{\mathbb{R}^d} \frac{\sigma^*(x)}{(2-\xi^*(x))(1-\xi^*(x))} P_X(\mathrm{d}x)$$

when $0 < \xi^*(x) < 1$ for all $x \in \mathbb{R}^d$. For a forecast in the GPD class, i.e. F_x is a GPD with shape parameter $\xi(x)$ and scale parameter $\sigma(x)$, then the risk $R_P(F)$ is equal to Bayes risk if and only if $\xi(x) = \xi^*(x)$ and $\sigma(x) = \sigma^*(x)$ P_X-a.e.

In the GPD regression framework, the conditions of the classes of distributions $\mathcal{D}^{(h,C,M)}$ can be interpreted as conditions on the parameters $\xi^*(x)$ and $\sigma^*(x)$. Condition *ii)* is equivalent to $\sigma^*(x) \leq M(2-\xi^*(x))(1-\xi^*(x))$ when $0 < \xi^*(x) < 1$, for all $x \in [0,1]^d$. The regularity condition *iii)* holds with constants C and h as soon as $x \mapsto \xi^*(x)$ and $x \mapsto \sigma^*(x)$ are both h-Hölder.

For example, the popular case were the shape parameter $\xi^*(x)$ and the scale parameter $\sigma^*(x)$ are assumed to be linearly dependent on x (i.e. $\xi^*(x) = \xi_0 + \xi_1 \cdot x$ and $\sigma^*(x) = \sigma_0 + \sigma_1 \cdot x$ with $\xi_1, \sigma_1 \in \mathbb{R}^d$) is in a class of distributions of Definition 2.2.

2.4 Conclusion and Discussion

We found that the optimal rate of convergence for distributional regression on $\mathcal{D}^{(h,C,M)}$ is of the same order as the optimal rate of convergence for point regression. Thus, with regard to the sample size n, distributional regression evaluated with the CRPS converges at the same rate as point regression even though the distributional estimate carries more information on the prediction of the underlying process.

We have also shown that the k-NN method and the kernel method reach this optimal rate of convergence, respectively in dimension $d \geq 2$ and in any dimension. However, these methods are not widely used in practice because of the limitations of their predictive power in moderate or high dimension $d \geq 3$ due to the curse of dimension. An extension of this work could be to study if state-of-the-art techniques reach the optimal rate of convergence obtained in this article. Random Forests (Breiman, 2001) methods, such as Quantile Regression Forests (Meinshausen, 2006) and Distributional Random Forests (Ćevid et al., 2022), appear to be natural candidates as they are based on a generalized notion of neighborhood and have been subject to recent development in weather forecast statistical postprocessing (see, e.g., Taillardat et al., 2016).

The results of this article were obtained for the CRPS, which is widely used in practice, but can easily be extended to the weighted CRPS in its standard uses. The weighted CRPS is defined as

$$\mathrm{wCRPS}(F, y) = \int_{\mathbb{R}} (F(z) - \mathbb{1}_{y \leq z})^2 w(z) \mathrm{d}z$$

with w the chosen weight. The weighted CRPS is used to put the focus of the score in specific regions of the outcome space (Gneiting and Ranjan, 2011). It is used in the study of extreme events by giving more weight to the extreme behavior of the distribution.

Moreover, an interesting development would be to obtain similar results for the rate of convergence with respect to different strictly proper scoring rules or metrics, for instance, energy scores or Wasserstein distances.

Acknowledgments

The authors acknowledge the support of the French Agence Nationale de la Recherche (ANR) under reference ANR-20-CE40-0025-01 (T-REX project) and of the Energy oriented Centre of Excellence-II (EoCoE-II), Grant Agreement 824158, funded within the Horizon2020 framework of the European Union. Part of this work was also supported by the ExtremesLearning grant from 80 PRIME CNRS-INSU and the ANR project Melody (ANR-19-CE46-0011).

2.5 Appendix

2.5.1 Proof of Proposition 2.1

For the simplicity of notation, we write simply $\mathbb{E}$ for the expectation with respect to $(X, Y) \sim P$ and $D_n \sim P^n$. The context makes it clear enough so as to avoid confusion.

Proof. Recall that for the CRPS, the excess risk is equal to

$$\mathbb{E}[R_P(\hat{F}_n)] - R_P(F^*) = \mathbb{E}\left[\int_{\mathbb{R}} |\hat{F}_{n,X}(z) - F^*_X(z)|^2 \mathrm{d}z\right]. \tag{2.15}$$

We first estimate $\mathbb{E}[|\hat{F}_{n,x}(z) - F^*_x(z)|^2]$ for fixed $x \in [0,1]^d$ and $z \in \mathbb{R}$. Denote by $X_{1:n}(x), \cdots, X_{k_n:n}(x)$ the nearest neighbors of x and by $Y_{1:n}(x), \ldots, Y_{k_n:n}(x)$ the associated values of the response variable. Conditionally on $X_{i:n}(x) = x_i$, $1 \leq i \leq k_n$, the random variables $Y_{i:n}(x)$, $1 \leq i \leq k_n$, are

independent and with distribution $F^*_{x_i}$, $1 \le i \le k_n$. This implies that, conditionally, $\hat{F}_{n,x}(z)$ is the average of the k_n independent random variables $\mathbb{1}_{\{Y_{i:n}(x) \le z\}}$ that have a Bernoulli distribution with parameter $F^*_{x_i}(z)$. Therefore, the conditional bias and variance are given by

$$\mathbb{E}[\hat{F}_{n,x}(z) - F^*_x(z) \mid X_i(x) = x_i, 1 \le i \le k_n] = \frac{1}{k_n} \sum_{i=1}^{k_n} \big(F^*_{x_i}(z) - F^*_x(z)\big)$$

$$\mathrm{Var}[\hat{F}_{n,x}(z) \mid X_i(x) = x_i, 1 \le i \le k_n] = \frac{1}{k_n^2} \sum_{i=1}^{k_n} F^*_{x_i}(z)(1 - F^*_{x_i}(z)).$$

Adding up the squared conditional bias and variance and integrating with respect to $X_{i:n}(x)$, $1 \le i \le k_n$, we obtain the mean squared error

$$\begin{aligned}
&\mathbb{E}\big[|\hat{F}_{n,x}(z) - F^*_x(z)|^2\big] \\
&= \mathbb{E}\Big[\Big(\frac{1}{k_n} \sum_{i=1}^{k_n} \big(F^*_{X_{i:n}(x)}(z) - F^*_x(z)\big)\Big)^2\Big] + \frac{1}{k_n^2} \sum_{i=1}^{k_n} \mathbb{E}\Big[F^*_{X_{i:n}(x)}(z)(1 - F^*_{X_{i:n}(x)}(z))\Big].
\end{aligned}$$

Using Jensen's inequality and integrating with respect to $P_X(\mathrm{d}x)\mathrm{d}z$, we deduce that the excess risk (2.15) satisfies

$$\begin{aligned}
\mathbb{E}[R_P(\hat{F}_n)] - R_P(F^*) \le \frac{1}{k_n} \sum_{i=1}^{k_n} \mathbb{E}\left[\int_{\mathbb{R}} (F^*_{X_{i:n}(X)}(z) - F^*_X(z))^2 \mathrm{d}z\right] \\
+ \frac{1}{k_n^2} \sum_{i=1}^{k_n} \mathbb{E}\left[\int_{\mathbb{R}} F^*_{X_{i:n}(X)}(z)(1 - F^*_{X_{i:n}(X)})\mathrm{d}z\right].
\end{aligned}$$

Using conditions *ii)* and *iii)* in the definition of the class $\mathcal{D}^{(h,C,M)}$ to bound from above the first and second term respectively, we get

$$\begin{aligned}
\mathbb{E}[R_P(\hat{F}_n)] - R_P(F^*) &\le \frac{C^2}{k_n} \sum_{i=1}^{k_n} \mathbb{E}\big[\|X_{i:n}(X) - X\|^{2h}\big] + \frac{M}{k_n} \\
&\le C^2 \mathbb{E}\big[\|X_{k_n:n}(X) - X\|^{2h}\big] + \frac{M}{k_n},
\end{aligned}$$

where the last inequality uses the fact that, by definition of nearest neighbors, the distances $\|X_{i:n}(X) - X\|$, $1 \le i \le k_n$, are non-increasing.

The last step of the proof is to use Theorem 2.4 from Biau and Devroye (2015) stating that

$$\mathbb{E}[\|X_{k_n:n}(X) - X\|^2] \le \begin{cases} 8\dfrac{k_n}{n} & \text{if } d = 1, \\ c_d \left(\dfrac{k_n}{n}\right)^{2/d} & \text{if } d \ge 2. \end{cases}$$

Together with the concavity inequality (as $h \in (0, 1]$)

$$\mathbb{E}[\|X_{k_n:n}(X) - X\|^{2h}] \le \mathbb{E}[\|X_{k_n:n}(X) - X\|^2]^h,$$

we deduce

$$\mathbb{E}[R_P(\hat{F}_n)] - R_P(F^*) \le \begin{cases} C^2 8^h \left(\dfrac{k_n}{n}\right)^h + \dfrac{M}{k_n} & \text{if } d = 1, \\ C^2 {c_d}^h \left(\dfrac{k_n}{n}\right)^{2h/d} + \dfrac{M}{k_n} & \text{if } d \ge 2, \end{cases}$$

concluding the proof of Proposition 2.1. □

2.5.2 Proof of Proposition 2.2

Proof. Equation (2.11) can be rewritten as

$$\hat{F}_{n,x}(z) = \frac{\sum_{i=1}^n \mathbb{1}_{\{X_i \in S_{x,h_n}\}} \mathbb{1}_{\{Y_i \leq z\}}}{nP_n(S_{x,h_n})},$$

with $S_{x,\epsilon}$ the closed ball centered at x of radius $\epsilon > 0$ and

$$P_n(\cdot) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{X_i \in \cdot\}}$$

the empirical measure corresponding to $X_1, \ldots, X_n$. Recall that we use the estimator $\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{Y_i \leq z\}}$ when $nP_n(S_{x,h_n}) = 0$.

Similarly as in the proof of the Proposition 2.1, a bias/variance decomposition of the squared error yields

$$\begin{aligned}
&\mathbb{E}\big[|\hat{F}_{n,x}(z) - F_x^*(z)|^2\big] \\
&= \mathbb{E}\left[\left(\frac{\sum_{i=1}^n \big(F^*_{X_i(x)}(z) - F_x^*(z)\big)\mathbb{1}_{\{X_i \in S_{x,h_n}\}}}{nP_n(S_{x,h_n})}\right)^2 \mathbb{1}_{\{nP_n(S_{x,h_n})>0\}}\right] \\
&\quad + \mathbb{E}\left[\frac{\sum_{i=1}^n F^*_{X_i}(z)(1 - F^*_{X_i}(z))\mathbb{1}_{\{X_i \in S_{x,h_n}\}}}{(nP_n(S_{x,h_n}))^2} \mathbb{1}_{\{nP_n(S_{x,h_n})>0\}}\right] \\
&\quad + \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{Y_i \leq z\}} - F_x^*(z)\right)^2 \mathbb{1}_{\{nP_n(S_{x,h_n})=0\}}\right] \\
&:= A_1(z) + A_2(z) + A_3(z).
\end{aligned}$$

The excess risk at $X = x$ is thus decomposed into three terms

$$\mathbb{E}\left[\int_{\mathbb{R}} |\hat{F}_{n,x}(z) - F_x^*(z)|^2 \mathrm{d}z\right] = \int_{\mathbb{R}} A_1(z)\mathrm{d}z + \int_{\mathbb{R}} A_2(z)\mathrm{d}z + \int_{\mathbb{R}} A_3(z)\mathrm{d}z$$

that we analyze successively.

The first term (bias) is bounded from above using Jensen's inequality and property *iii)* of $\mathcal{D}^{(h,C,M)}$:

$$\begin{aligned}
\int_{\mathbb{R}} A_1(z)\mathrm{d}z &\leq \mathbb{E}\left[\frac{\sum_{i=1}^n \int_{\mathbb{R}} \big(F^*_{X_i(x)}(z) - F_x^*(z)\big)^2 \mathrm{d}z \mathbb{1}_{\{X_i \in S_{x,h_n}\}}}{nP_n(S_{x,h_n})} \mathbb{1}_{\{nP_n(S_{x,h_n})>0\}}\right] \\
&\leq \mathbb{E}\left[\frac{\sum_{i=1}^n C^2 \|X_i - x\|^{2h} \mathbb{1}_{\{X_i \in S_{x,h_n}\}}}{nP_n(S_{x,h_n})} \mathbb{1}_{\{nP_n(S_{x,h_n})>0\}}\right] \\
&\leq C^2 {h_n}^{2h}.
\end{aligned}$$

The second term (variance) is bounded using property *ii)* of $\mathcal{D}^{(h,C,M)}$ and an elementary result for the binomial distribution:

$$\begin{aligned}
\int_{\mathbb{R}} A_2(z)\mathrm{d}z &= \mathbb{E}\left[\frac{\sum_{i=1}^n \int_{\mathbb{R}} F^*_{X_i}(z)(1 - F^*_{X_i}(z))\mathrm{d}z \mathbb{1}_{\{X_i \in S_{x,h_n}\}}}{(nP_n(S_{x,h_n}))^2} \mathbb{1}_{\{nP_n(S_{x,h_n})>0\}}\right] \\
&\leq M\mathbb{E}\left[\frac{\mathbb{1}_{\{nP_n(S_{x,h_n})>0\}}}{nP_n(S_{x,h_n})}\right] \\
&\leq \frac{2M}{nP_X(S_{x,h_n})}.
\end{aligned}$$

In the last line, we use that $Z = nP_n(S_{x,h_n})$ follows a binomial distribution with parameters n and $p = P_X(S_{x,h_n})$ so that $\mathbb{E}\left[\frac{1}{Z}\mathbb{1}_{\{Z>0\}}\right] \leq \frac{2}{(n+1)p}$, see Lemma 4.1 in Györfi et al. (2002).

The last term is a remainder term and is bounded by

$$\begin{aligned}\int_{\mathbb{R}} A_3(z)\mathrm{d}z \leq \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\int_{\mathbb{R}}\left(F^*_{X_i}(z) - F^*_x(z)\right)^2 \mathrm{d}z\mathbb{1}_{\{nP_n(S_{x,h_n})=0\}}\right] \\ + \mathbb{E}\left[\frac{1}{n^2}\sum_{i=1}^{n}\int_{\mathbb{R}} F^*_{X_i}(z)(1-F^*_{X_i}(z))\mathrm{d}z\mathbb{1}_{\{nP_n(S_{x,h_n})=0\}}\right].\end{aligned}$$

Properties *ii)* and *iii)* of $\mathcal{D}^{(h,C,M)}$ and the fact that $\|X_i - x\| \leq \sqrt{d}$ imply

$$\begin{aligned}\int_{\mathbb{R}} A_3(z)\mathrm{d}z &\leq \left(C^2 d^h + \frac{M}{n}\right)\mathbb{E}\left[\mathbb{1}_{\{nP_n(S_{x,h_n})=0\}}\right] \\ &\leq \left(C^2 d^h + \frac{M}{n}\right)e^{-nP_X(S_{x,h_n})}.\end{aligned}$$

For the second inequality, we use that $\mathbb{P}(Z=0) = (1-p)^n \leq e^{-np}$ where $Z = nP_n(S_{x,h_n})$ follows a binomial distribution with parameters n and $p = P_X(S_{x,h_n})$.

Collecting the three terms, we obtain the following upper bound for the excess risk at $X = x$:

$$\mathbb{E}\left[\int_{\mathbb{R}} |\hat{F}_{n,x}(z) - F^*_x(z)|^2 \mathrm{d}z\right] \leq C^2 {h_n}^{2h} + \frac{2M}{nP_X(S_{x,h_n})} + \left(C^2 d^h + \frac{M}{n}\right)e^{-nP_X(S_{x,h_n})}.$$

We finally integrate this bound with respect to $P_X(\mathrm{d}x)$. According to Equation (5.1) in Györfi et al. (2002), there exists a constant $\tilde{c}_d$ depending only on d such that

$$\int_{[0,1]^d} \frac{1}{nP_X(S_{x,h_n})} P_X(\mathrm{d}x) \leq \frac{\tilde{c}_d}{nh_n^d}.$$

Note that $\tilde{c}_d$ can be chosen as $\tilde{c}_d = d^{d/2}$. We also have

$$\begin{aligned}\int_{[0,1]^d} e^{-nP_X(S_{x,h_n})} P_X(\mathrm{d}x) &\leq \max_{u\geq 0} ue^{-u} \int_{[0,1]^d} \frac{1}{nP_X(S_{x,h_n})} P_X(\mathrm{d}x) \\ &\leq \frac{\tilde{c}_d}{nh_n^d}.\end{aligned}$$

We obtain thus

$$\begin{aligned}\mathbb{E}[R_P(\hat{F}_n)] - R_P(F^*) &= \mathbb{E}\left[\int_{\mathbb{R}} |\hat{F}_{n,x}(z) - F^*_x(z)|^2 \mathrm{d}z\right] \\ &\leq C^2 {h_n}^{2h} + \tilde{c}_d \frac{2M + C^2 d^h + \frac{M}{n}}{n{h_n}^d}.\end{aligned}$$

□

2.5.3 Proof of Proposition 2.3

The proof of Proposition 2.3 relies on the next two elementary lemmas. The first one states that for a binary outcome $Y \in \{0, L\}$, forecasters should focus on binary forecast $F \in \mathcal{M}(\{0, L\})$ only, which is very natural. More precisely, any predictive distribution $F \in \mathcal{M}(\mathbb{R})$ can be associated with $F \in \mathcal{M}(\{0, L\})$ with a better expected CRPS.

Lemma 2.1. *Let $G \in \mathcal{M}(\{0, L\})$. For $F \in \mathcal{M}(\mathbb{R})$, the distribution*

$$\tilde{F}(z) = (1-m)\mathbb{1}_{0\leq z} + m\mathbb{1}_{L\leq z} \text{ with } m = \frac{1}{L}\int_0^L (1-F(z))\mathrm{d}z$$

satisfies

$$\overline{\mathrm{CRPS}}(\tilde{F}, G) \leq \overline{\mathrm{CRPS}}(F, G).$$

Proof. Let $F \in \mathcal{M}(\mathbb{R})$ and $G \in \mathcal{M}(\{0, L\})$. We have

$$\begin{aligned}\overline{\mathrm{CRPS}}(F, G) &= \int_{\mathbb{R}}\int_{\mathbb{R}} (F(z) - \mathbb{1}_{y\leq z})^2 \mathrm{d}z G(\mathrm{d}y) \\ &\geq \int_{\mathbb{R}}\int_0^L (F(z) - \mathbb{1}_{y\leq z})^2 \mathrm{d}z G(\mathrm{d}y)\end{aligned}$$

Because $1-m$ is the mean value of F on $[0, L]$, we have for $y \in \{0, L\}$

$$\int_0^L (F(z) - \mathbb{1}_{y\leq z})^2 \mathrm{d}z \geq \int_0^L ((1-m) - \mathbb{1}_{y\leq z})^2 \mathrm{d}z.$$

Integrating with respect to $G(\mathrm{d}y)$, we deduce

$$\overline{\mathrm{CRPS}}(F, G) \geq \int_{\mathbb{R}}\int_0^L ((1-m) - \mathbb{1}_{y\leq z})^2 \mathrm{d}z G(\mathrm{d}y).$$

The right-hand side equals $\overline{\mathrm{CRPS}}(\tilde{F}, G)$ and we conclude

$$\overline{\mathrm{CRPS}}(F, G) \geq \overline{\mathrm{CRPS}}(\tilde{F}, G).$$

□

Lemma 2.2 shows that for binary outcome and predictions, the CRPS reduces to a quantity proportional to the Brier score (Brier, 1950)

$$\mathrm{Brier}(p, y) = (y-p)^2, \quad y \in \{0, 1\}, p \in [0, 1],$$

which is closely related to the mean squared error used in regression.

Lemma 2.2. *For all $y \in \{0, L\}$ and $F(z) = (1-p)\mathbb{1}_{0\leq z} + p\mathbb{1}_{L\leq z} \in \mathcal{M}(\{0, L\})$ with $p \in [0, 1]$, it holds*

$$\mathrm{CRPS}(F, y) = L\mathrm{Brier}(p, \frac{y}{L}) = L(\frac{y}{L} - p)^2.$$

Proof. We compute

$$\begin{aligned}\mathrm{CRPS}(F, y) &= \int_0^L (1 - p - \mathbb{1}_{y\leq z})^2 \mathrm{d}z \\ &= \begin{cases} Lp^2 & \text{if y=0} \\ L(1-p)^2 & \text{if y=L} \end{cases}.\end{aligned}$$

In both cases, this equals $L(\frac{y}{L} - p)^2 = L\mathrm{Brier}(p, \frac{y}{L})$. □

Proof of Proposition 2.3. Since only binary outcomes are considered in the class $\mathcal{B}^{(h,C,L)}$, Lemma 2.1 implies that

$$\inf_{\hat{F}_n} \sup_{P\in\mathcal{B}^{(h,C,L)}} \left\{\mathbb{E}[R_P(\hat{F}_n)] - R_P(F^*)\right\} = \inf_{\tilde{F}_n} \sup_{P\in\mathcal{B}^{(h,C,L)}} \left\{\mathbb{E}[R_P(\tilde{F}_n)] - R_P(F^*)\right\}$$

where the infimum are taken over models $\hat{F}_n$ and $\tilde{F}_n$ trained on the first observations $(X_i, Y_i)_{1\leq i\leq n}$ and with values in $\mathcal{M}(\mathbb{R})$ and $\mathcal{M}(\{0, L\})$, respectively. Indeed, the left-hand side is a priori smaller since the family $\hat{F}_n$ is larger but Lemma 2.1 ensures that each model $\hat{F}_n$ can be associated with a model $\tilde{F}_n$ with equal or lower expected score.

We then apply Lemma 2.2. For a binary outcome, the conditional distribution of Y given $X = x$ writes

$$F_x^*(z) = (1 - m(x))\mathbb{1}_{0\leq z} + m(x)\mathbb{1}_{L\leq z},$$

and the model $\tilde{F}_n$ with values in $\mathcal{M}(\{0, L\})$ takes the form

$$\tilde{F}_{n,x}(z) = (1 - m_n(x))\mathbb{1}_{0\leq z} + m_n(x)\mathbb{1}_{L\leq z},$$

with $m(x) = \frac{1}{L}\int_0^L (1 - F_x^*(z))\mathrm{d}z$ and $m_n(x) = \frac{1}{L}\int_0^L (1 - \hat{F}_{n,x}(z))\mathrm{d}z$.
Then Lemma 2.2 implies

$$\begin{aligned}
\mathbb{E}[R_P(\hat{F}_n)] - R_P(F^*) &= \mathbb{E}\left[\mathrm{CRPS}(\hat{F}_{n,X}, Y) - \mathrm{CRPS}(F_X^*, Y)\right] \\
&= L\mathbb{E}\left[(Y/L - m_n(X))^2 - (Y/L - m(X))^2\right] \\
&= L\mathbb{E}\left[(m_n(X) - m(X))^2\right],
\end{aligned}$$

which corresponds to the excess risk in regression with squared error loss. The property *iii*) of $\mathcal{B}^{(h,C,L)}$ is equivalent to

$$|m(x) - m(x')|^h \leq C\|x - x'\|^h, \quad x \in [0, 1]^d,$$

which is the standard regularity assumption on the regression function m. Using the result of Problem 3.3 in Györfi et al. (2002) dealing with binary models, we finally obtain that the sequence $a_n = n^{-\frac{2h}{2h+d}}$ is a lower minimax rate of convergence for this class of distributions and more precisely that Equation (2.12) holds. □

Chapter 3

Distributional regression U-Nets for the postprocessing of precipitation ensemble forecasts

This chapter reproduces an article submitted to *Artificial Intelligence for the Earth Systems*, and written by Romain Pic[1], Clément Dombry[1], Philippe Naveau[2] and Maxime Taillardat[3].

Abstract Accurate precipitation forecasts have a high socio-economic value due to their role in decision-making in various fields such as transport networks and farming. We propose a global statistical postprocessing method for grid-based precipitation ensemble forecasts. This U-Net-based distributional regression method predicts marginal distributions in the form of parametric distributions inferred by scoring rule minimization. Distributional regression U-Nets are compared to state-of-the-art postprocessing methods for daily 21-h forecasts of 3-h accumulated precipitation over the South of France. Training data comes from the Météo-France weather model AROME-EPS and spans 3 years. A practical challenge appears when consistent data or reforecasts are not available.

Distributional regression U-Nets compete favorably with the raw ensemble. In terms of continuous ranked probability score, they reach a performance comparable to quantile regression forests (QRF). However, they are unable to provide calibrated forecasts in areas associated with high climatological precipitation. In terms of predictive power for heavy precipitation events, they outperform both QRF and semi-parametric QRF with tail extensions.

Contents

[1]Université de Franche Comté, CNRS, LmB (UMR 6623), F-25000 Besançon, France
[2]Laboratoire des Sciences du Climat et de l'Environnement, UMR 8212, CEA-CNRS-UVSQ, EstimR, IPSL & U Paris-Saclay, Gif-sur-Yvette, France
[3]CNRM, Université de Toulouse, Météo-France, CNRS, Toulouse, France

3.1 Introduction

Correctly forecasting precipitation is crucial for decision-making in various fields such as flood levels, transport networks, water resources and farming, among others (see, e.g., Olson et al. 1995). Moreover, high-impact events are expected to intensify in the future as a consequence of climate change (Planton et al., 2008). Numerical weather prediction (NWP) systems have been continuously improving to take into account uncertainty of the atmosphere and the limitations of their physical modeling (Bauer et al., 2015). NWP systems produce ensemble forecasts, consisting of multiple runs of deterministic scenarios with different parameters. Nonetheless, raw ensemble forecasts suffer from bias and underdispersion (see, e.g., Hamill and Colucci 1997; Bauer et al. 2015; Ben Bouallègue et al. 2016; Baran and Nemoda 2016). This phenomenon affects all NWP systems regardless of the weather service and of the variable of interest. Furthermore, the limited number of ensemble members coupled with underdispersion implies that raw ensemble forecasts may have a limited predictive power regarding extremes (Williams et al., 2013). In order to correct these systematic errors, it has become standard practice to use statistical postprocessing of ensemble prediction systems (EPS) in both research and operations.

A popular spatial statistical postprocessing strategy consists of separately postprocessing marginal distributions at each location and the spatial dependence structure. Numerous methods for postprocessing univariate marginals have been developed over the past two decades. There has been a rise in the number of machine learning based statistical postprocessing techniques as they provide a flexible framework enabling the modeling of complex relationships between the output of NWP models and the target variable. Moreover, they facilitate the use of a large number of predictors. These methods range from well-established statistical learning techniques, such as random forests (Taillardat et al., 2016) or gradient boosting (Messner et al., 2017), to neural networks or deep learning techniques, such as fully connected neural networks (Rasp and Lerch, 2018) and transformers (Ben Bouallègue et al., 2024b). For a thorough review of the existing statistical postprocessing techniques, readers may refer to Vannitsem et al. (2021) and Schulz and Lerch (2022b). Once calibrated univariate marginals are obtained, the spatial dependence structure may be needed by downstream applications. The spatial dependence structure can be obtained from the raw ensemble as done by ensemble copula coupling (ECC; Schefzik et al. 2013) and its variants (e.g., Ben Bouallègue et al. 2016) or from historical observations as done by Schaake shuffle (ScS; Clark et al. 2004). Alternatively, if raw ensembles or historical data do not model the spatial dependence sufficiently well, it can be postprocessed using adapted techniques (see, e.g., Schefzik and Möller 2018).

An alternative postprocessing strategy consists of direct postprocessing of raw ensemble members to obtain calibrated members. This can be achieved by postprocessing each member individually (Van Schaeybroeck and Vannitsem, 2015) or by using ensemble-agnostic methods (Ben Bouallègue et al., 2024b).

In order to circumvent (potential) data scarcity, it is common to use parametric methods as they are usually less affected by smaller training datasets. The choice of a specific parametric

distribution can be motivated by prior knowledge (or assumption) on the distribution of the variable of interest. Parametric methods can enable extrapolation beyond the range available in the training data, which is of interest to consider extreme events (see, e.g., Friederichs et al. 2018 and Taillardat et al. 2019). In particular, certain meteorological variables have a heavy-tailed distribution; thus, a parametric method can be used to ensure that postprocessed distributions will have an appropriate tail behavior (e.g., Lerch and Thorarinsdottir 2013).

Previous studies, such as Hemri et al. (2014) and Taillardat and Mestre (2020), have highlighted that all meteorological quantities do not represent the same difficulty in terms of postprocessing. Variables with heavy-tailed climatological distributions or variables with short-scale spatio-temporal dependence (e.g., rainfall or wind gusts) are more difficult to treat than light-tailed variables or spatially smooth variables (e.g., surface temperature or sea level pressure). In the same vein, Schulz and Lerch (2022b) states that "wind gusts are a challenging meteorological target variable as they are driven by small-scale processes and local occurrence, so that their predictability is limited even for numerical weather prediction (NWP) models run at convection-permitting resolutions."

NWP models produce forecasts on a grid that are of interest to downstream applications (Hamill, 2018, Section 7.3.2). However, consistent gridded data suited to postprocessing is computationally costly since reanalyses and reforecasts of gridded products are demanding in terms of both storage and computation. Numerous observation networks are station-based (e.g., temperature, wind speed, or pressure), but they vary in coverage and quality. When forecasts are required at nearby locations, spatial modeling procedures are required. Both station-based and grid-based approaches present benefits and drawbacks (Hamill, 2018, Section 7.3.2). No preference has reached a consensus for any variable, but Feldmann et al. (2019) shows that the relative improvement is greater for station-based 2-m temperature postprocessing when station-based observations are used. In the case of precipitation, observations can be measured by hybrid observations (gauge-adjusted radar images), allowing for improvement in the quality of gridded postprocessing.

As mentioned by Schulz and Lerch (2022b), one of the main challenges of postprocessing is to preserve the spatio-temporal information while optimally utilizing the whole available input data. This motivates the use of global statistical postprocessing models (e.g., a single model for multiple locations). Distributional regression networks (DRN; Rasp and Lerch 2018) use an embedding module to learn a representation of stations, allowing the model to learn from nearby and similar stations in order to preserve the spatial information of the data. When working with gridded data, a postprocessing method could benefit from taking into account this spatial structure of the data within its architecture. Convolutional neural networks (CNN) rely on the image-like structure of their input. Numerous CNN-based methods have been developed to perform postprocessing (see, e.g., Dai and Hemri 2021 and Lerch and Polsterer 2022). Here, we want the output of the statistical postprocessing method to be grid-based. U-Net (Ronneberger et al., 2015) architectures appear to be a natural solution to preserve the spatial structure of the data. U-Nets use a sequence of convolutional blocks to learn complex features and upscaling blocks to retrieve parameters of interest at the desired resolution. We propose a U-Net-based method to postprocess marginals at each grid point using predictors at nearby grid points for high-resolution precipitation ensemble forecasts.

The paper is organized as follows. Section 3.2 presents the dataset used in this study. In Section 3.3, three state-of-the-art methods composing the reference methods of this study, namely quantile regression forests (QRF), QRF with tail extension (TQRF) and DRN, are presented and compared based on their known benefits and limitations. A U-Net-based method, called distributional regression U-Nets (DRU), is introduced and compared with U-Net-based postprocessing methods in the literature. The predictive performance of the models is compared in

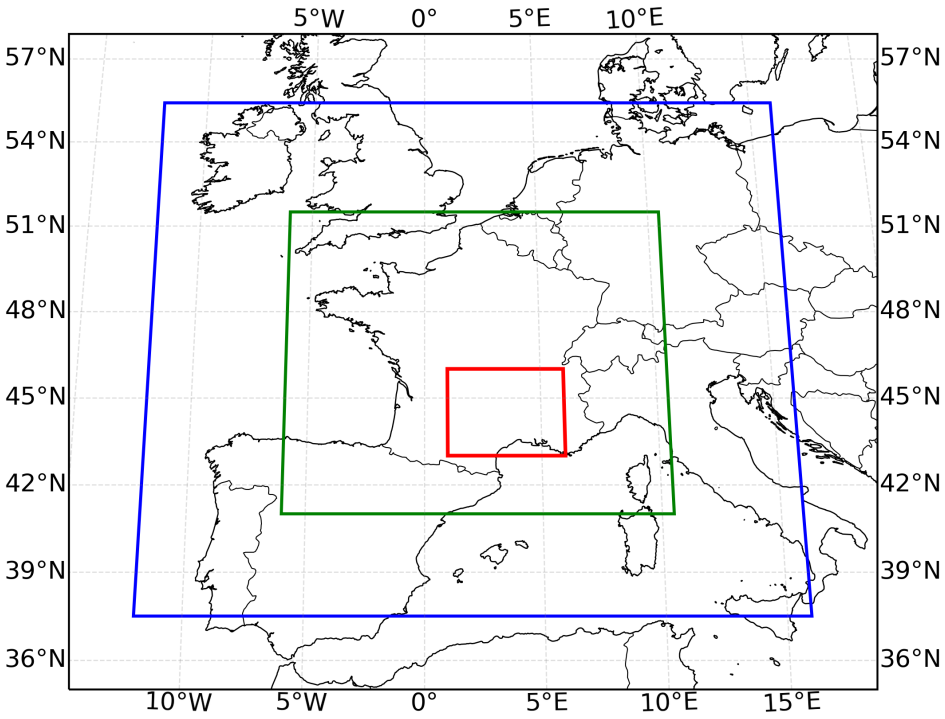


Figure 3.1: Domains covered by AROME-EPS (blue), ANTILOPE (green) and the region of interest (red).

terms of multiple univariate metrics in Section 3.4. An emphasis is put on the predictive performance of extremes. Finally, Section 3.5 sums up the performance of DRU and offers possible perspectives.

The code used to implement the different methods and their verification is publicly available[4].

3.2 Data

In this study, we focus on 3-h accumulated precipitation over the South of France (see Fig. 3.1) at a forecast lead time of 21-h initialized at 15:00UTC daily. Ensemble forecasts are taken from the 17-member limited area ensemble forecasting system AROME-EPS (Bouttier et al., 2015) driven by a subsampling of the global[5] PEARP ensemble. AROME-EPS produces ensembles with one control member and 16 perturbed members for forecasts up to 51 hours on four different initialization times. It produces a gridded ensemble over Western Europe with a horizontal resolution of 0.025° based on a model run at 1.3 km resolution. The probabilistic forecasts are compared to 3-h accumulated precipitation data obtained from the gauge-adjusted radar product ANTILOPE (Champeaux et al., 2009), which has a spatial resolution of 0.001° over Western Europe. We project observations of ANTILOPE onto the AROME-EPS grid using bilinear interpolation.

[4] https://github.com/pic-romain/unet-pp
[5] in the sense of globe-wide

The region of interest in this study covers areas, such as the *Cévennes*, prone to heavy precipitation events (HPEs) (Ricard et al., 2012). HPEs affect Mediterranean coastal regions regularly causing flash floods. Mediterranean HPEs are typically characterized by quasi-stationary convective precipitation and may have limited predictability due to their intensity and being very local (Caumont et al., 2021). Statistical postprocessing methods can help improve forecasting such events.

Our period of interest spans 4 years from November 2019 to October 2023. The period from November 2019 to October 2022 is used as a training/validation dataset using 7-fold cross-validation to tune hyperparameters of the models. The folds are based on the day of the week. The period from November 2022 to October 2023 is used as a hold-out test set. All the results of Section 3.4 are provided for models trained on the entirety of the training/validation dataset and evaluated on the test dataset. The dataset is composed of forecasts and reforecasts from two different cycles of AROME-EPS. Consistency of both raw ensembles and observations is important since independent and identically distributed (i.i.d.) data is assumed. The two cycles of AROME-EPS used, namely 43t2 and 46t1, only have minor differences, making the i.i.d. assumption reasonable.

We use summary statistics of the AROME-EPS ensemble as predictors. The following variables were selected based on experts' opinions: precipitation, convective available potential energy, maximal reflectivity, pseudo wet-bulb potential temperature, relative humidity and AROME convection index. For each of these variables, the mean, the minimum, the maximum and the standard deviation of the raw ensemble were computed at each grid point and used as predictors.

In addition to summary statistics from AROME-EPS, distributional regression U-Nets (DRU) use constant fields carrying information about the topography and the type of terrain as predictors. The constant fields used are the altitude, a land-sea mask, the distance to sea and the first four components of a principal component analysis decomposition called AURHELY (Bénichou, 1994). Lerch and Polsterer (2022) showcased that the use of constant fields, such as altitude or orography, improves the performance of DRN. The first four components of AURHELY can be interpreted as local peak/depression, Northern/Southern slope, Eastern/Western slope and saddle effects, respectively. Figure 3.2 shows the seven constant fields used as predictors in DRU. Table 3.1 summarizes the predictors issued from both the raw ensemble and constant fields. Table 3.2 lists the dimensions of the dataset.

Type	Variable
Raw ensemble (mean, min, max, sd)	Precipitation
	Convective available potential energy
	Maximal reflectivity
	Pseudo wet-bulb potential temperature
	Relative humidity
	AROME convection index
Constant fields	Altitude
	Land-sea mask
	AURHELY components (1-4)
	Distance to sea

Table 3.1: List of weather and topographic variables used as predictors.

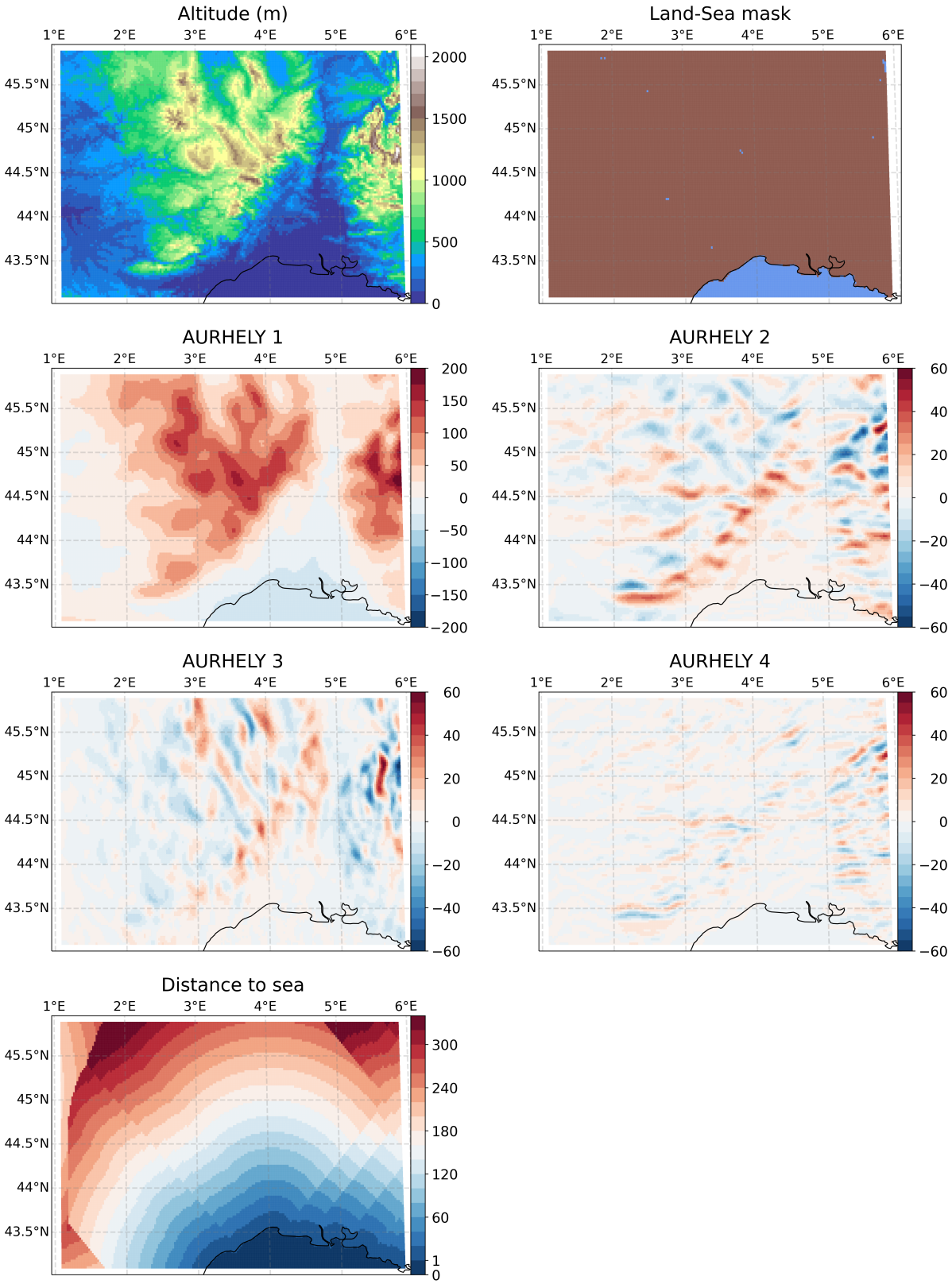


Figure 3.2: Constants fields used as predictors in distributional regression U-Nets: altitude, land-sea mask, four first components of AURHELY procedure, and distance to the sea.

Variable	Value	Description
d	31	number of predictors
H	112	height (in grid points) of the region of interest (latitude)
W	192	width (in grid points) of the region of interest (longitude)
$n_{trainval}$	1091	# of days in the training/validation dataset
n_{test}	365	# of days in the test dataset

Table 3.2: Dimensions of the dataset used in this study.

3.3 Methods

We compare several postprocessing methods for the marginal distributions of gridded spatial ensemble forecasts of 3-h accumulated precipitation over the South of France. In a complete postprocessing scheme used operationally, the multivariate dependencies can then be retrieved using ECC or ScS, for example. We compare our U-Net-based distributional regression method to two benchmark methods: quantile regression forest (QRF; Taillardat et al. 2016) and QRF with tail extension (TQRF; Taillardat et al. 2019). The performance of postprocessed forecasts using these different methods will be compared to the performance of the raw ensemble. Additionally, we recall distributional regression networks (DRN; Rasp and Lerch 2018) since our method can be seen as an extension of this approach.

These methods differ in their degree of reliance on parametric distributions (nonparametric, semi-parametric and parametric), in the fact of being local (i.e., a different model for each grid point) or global (i.e., a single model for the whole grid). Among global methods, differences lie in the representation of the spatial structure of the data. We briefly present the benchmark techniques and their limitations.

3.3.1 Quantile regression forests (QRF)

Quantile regression forests (QRF; Meinshausen 2006) is a nonparametric method able to predict conditional quantiles or, more generally, a conditional distribution. The method is based on random forests (Breiman, 2001). Similarly, it uses the data in terminal nodes (i.e., leaves) to compute a weighted average of empirical distributions. QRFs have proven their performance for postprocessing of wind speed and temperature forecasts (Taillardat et al., 2016) and for precipitation forecasts (Whan and Schmeits, 2018; van Straaten et al., 2018). QRFs can outperform complex postprocessing methods, such as neural network (NN-)based methods, at specific locations due to their local adaptability (Rasp and Lerch, 2018; Schulz and Lerch, 2022b). Moreover, QRF is used operationally as a postprocessing method at Météo-France (Taillardat and Mestre, 2020). This, as well as its overall performance, makes it a relevant benchmark method for this study.

QRFs are known to have three main limitations: potential spatial inconsistency, storage memory voracity (Taillardat and Mestre, 2020) and inability to extrapolate. The fact that QRF is a local model (i.e., a different model is used for each location, lead time, and variable) may cause problems. There is no guarantee that the output of the models is consistent spatially or temporally. Additionally, QRFs need to store the construction parameters (such as variables and thresholds of splits) of each tree of the forest and the samples used for training. This latter limitation results in the need to store a large number of parameters (especially when working

with gridded data) to perform postprocessing. Lastly, QRF is incapable of extrapolating as its output is a weighted average of the training samples and does not provide a model for the distribution tail.

3.3.2 Quantile regression forest with tail extension (TQRF)

In order to circumvent the extrapolation inability of QRF, semi-parametric methods based on a combination of parametric modeling and random forest were proposed. Schlosser et al. (2019) introduced distributional regression forests using maximum likelihood to infer the parameters of a censored Gaussian distribution. Taillardat et al. (2019) proposed a method using probability-weighted moments (Diebolt et al., 2007) on the output of QRF to infer the parameters of an extended generalized Pareto distribution (EGPD; Naveau et al. 2016). The EGPD is a flexible parametric class of distributions able to jointly model the whole range of the distribution while in alignment with extreme value theory, without the requirement of threshold selection. The methods proposed in Schlosser et al. (2019), Taillardat et al. (2019) and, more recently, Muschinski et al. (2023) can all be adapted to any suitable parametric distribution. We choose to use the semi-parametric method of Taillardat et al. (2019) based on probability-weighted moments inference.

Our implementation of TQRF differs from the original method described in Taillardat et al. (2019). It uses refinements that have proven to be useful in operational settings: the tail extension is only activated if the QRF forecast assigns a large enough probability of exceedance of certain levels of interest, and in that case, only the quantiles that are higher for the fitted distribution than in the output of the QRF are updated. Moreover, we did not use EGPD because, while the QRF+EGPD is robust and efficient, the minimization of its continuous ranked probability score (CRPS; Matheson and Winkler 1976) for parameter inference is not direct due to its complex form (Taillardat et al., 2019, 2022). These implementation issues could, for example, be circumvented by using Monte-Carlo sampling to estimate the CRPS or by fixing the tail parameter to its climatological value.

Instead of the EGPD, the generalized truncated/censored normal distribution (GTCND; Jordan et al. 2019) and the censored-shifted gamma distribution (CSGD; Scheuerer and Hamill 2015a) are used as tail extensions of the QRF and as parametric distributions for DRU. The GTCND used here has a lower endpoint equal to 0 and no upper endpoint and its cumulative distribution function (cdf) is defined as

$$F^{\text{gtcnd}}_{L,\mu,\sigma}(z) = \begin{cases} L + \frac{1-L}{1-\Phi(-\mu/\sigma)}\left(\Phi(\frac{z-\mu}{\sigma}) - \Phi(-\mu/\sigma)\right) & \text{if } z \geq 0 \\ 0 & \text{if } z < 0 \end{cases},$$

where $0 \leq L \leq 1$ is the probability of a dry event (i.e., absence of precipitation), Φ is the cdf of the standard normal distribution, $\mu \in \mathbb{R}$ is the location parameter of the truncated normal distribution and $\sigma > 0$ is its scale parameter. The cdf of the CSGD is defined as

$$F^{\text{csgd}}_{k,\theta,\delta}(z) = \begin{cases} G_k(\frac{z-\delta}{\theta}) & \text{if } z \geq 0 \\ 0 & \text{if } z < 0 \end{cases},$$

where G_k is the cdf of the gamma distribution of shape $k > 0$, θ is the scale parameter and $\delta < 0$ is a shift parameter. The probability of dry events has a point mass of $G_k(-\delta/\theta)$. These distributions are both suited to the forecast of precipitation since they have point masses in 0 and take positive values. Moreover, the CSGD can reflect the variations of skewness observed in precipitation distributions (Scheuerer and Hamill, 2015a). Details on the moments method

for GTCND and CSGD, as well as CRPS formulas, are provided in Appendix 3.6.1 and Appendix 3.6.2.

We denote QRF+*distrib* the TQRF method where *distrib* is the name of the parametric distribution family. The QRF+EGPD method is used operationally for rainfall postprocessing at Météo-France (Taillardat and Mestre, 2020). Nonetheless, this semi-parametric method remains local and thus also suffers from both potential spatial inconsistency and memory voracity (Taillardat and Mestre, 2020). To bypass these limitations, methods need to be global (i.e., use one model for all locations) while staying efficient locally.

3.3.3 Distributional regression networks (DRN)

Rasp and Lerch (2018) proposed distributional regression networks (DRN), a NN-based approach to postprocess 2-m temperature forecasts. DRN is a global model predicting the parameters of a distribution of interest. It leverages the flexibility of NN to model the dependency of parameters on the covariables (used as input of DRN). DRN can be seen as an extension of EMOS (Gneiting et al., 2005), which itself fits a parametric distribution where the parameters linearly depend on summary statistics of the raw ensemble. DRN is a global model thanks to the presence of an embedding module within its architecture, allowing the network to learn location-specific parameters and to benefit from data at similar locations. DRN learns the embedding and parameters of a dense NN by minimizing a strictly proper scoring rule (Gneiting and Katzfuss, 2014) such as the CRPS.

Rasp and Lerch (2018) and Schulz and Lerch (2022b) have shown that DRN outperforms other state-of-the-art methods in most stations over Germany for the postprocessing of temperature and wind gusts, respectively. Moreover, Schulz and Lerch (2022b) studied other NN-based postprocessing techniques, namely Bernstein quantile network (BQN; Bremnes 2020) and histogram estimation network (HEN; see, e.g., Scheuerer et al. 2020 and Veldkamp et al. 2021). BQN and HEN are nonparametric approaches where NNs learn the coefficient of Bernstein polynomials to predict a quantile function and probabilities of bins to predict a probability density function (pdf), respectively. At particular stations, BQN outperforms other postprocessing techniques, including DRN, for wind gust forecasts.

In spite of being a global model, the architecture of DRN makes it ill-suited to gridded data. Its architecture does not use knowledge of the spatial structure of the points and thus has to try to learn it through its embedding module. Moreover, DRN only uses information available at the location of interest as predictors. Convolutional neural network (CNN)-based architectures make use of the gridded structure of the data and can use the information at neighboring locations as a predictor. Lerch and Polsterer (2022) studied a modified DRN architecture using the representation of global fields from a convolutional auto-encoder as predictors and showed an improvement in skill compared to regular DRN.

DRNs' architecture makes their implementation on gridded data very costly. They need to flatten the data across locations (i.e., reshape it into a 1D vector), and they cannot benefit from GPU computing. For these reasons and their impact on the search for optimal hyperparameters, DRNs are not used as a benchmark method in this study.

3.3.4 Distributional regression U-Nets (DRU)

Convolutional blocks are the main ingredient of CNN-based architectures. The simplest convolutional blocks are composed of a convolutional layer and a max-pooling layer. The role of the convolutional layer is to learn kernels able to extract useful features from the input of the convolutional block. The max-pooling layer reduces the resolution of the features, allowing the

following layers to work at broader scales. The succession of convolutional blocks allows CNNs to learn patterns at different spatial scales and to learn complex patterns (see, e.g., Simonyan and Zisserman 2015). CNN-based architectures have been used in numerous postprocessing studies (e.g., Dai and Hemri 2021; Veldkamp et al. 2021; Li et al. 2022; Chapman et al. 2022; Lerch and Polsterer 2022).

Since we are interested in global models using the data's gridded structure and want the output to be the distributional parameters of marginals on the same grid, we use a U-Net architecture (Ronneberger et al., 2015). The U-Net architecture was initially designed for images but is compatible with gridded data to obtain a grid-based output. It has been used for various postprocessing applications. Grönquist et al. (2021) used it in a bias/uncertainty postprocessing scheme of temperature and geopotential forecasts. Dai and Hemri (2021) used a U-Net as a generator within a conditional generative adversarial network (cGAN) for cloud cover postprocessing. Hu et al. (2023) used U-Nets to predict the parameters of a CSGD corresponding to the postprocessed daily precipitation given a deterministic forecast. Horat and Lerch (2024) used U-Nets to perform postprocessing of temperature and precipitation at the sub-seasonal to seasonal scale. The task is a three-level classification problem with below-normal, near-normal and above-normal conditions as classes. Ben Bouallègue et al. (2024b) used transformers within a U-Net architecture to postprocess ensemble members directly with temperature and precipitation as variables of interest.

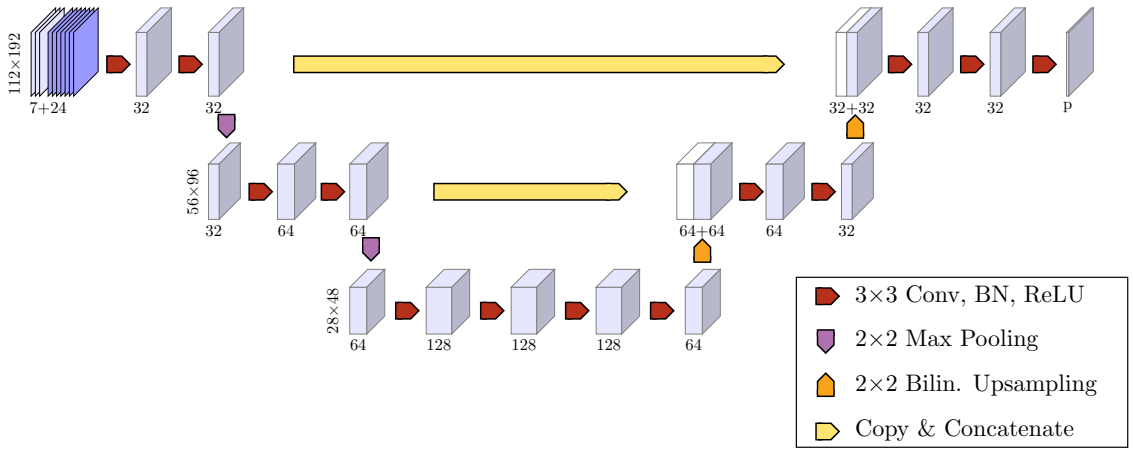


Figure 3.3: Architecture of distributional regression U-Nets. *Conv* stands for convolution, *BN* stands for batch normalization, *ReLU* stands for rectified linear unit and *Bilin. Upsampling* stands for bilinear upsampling. p is the number of distribution parameters: for GTCND and CSGD, $p = 3$.

The U-Net architecture used in this work is presented in Figure 3.3. The U-Net input is a concatenation of constant fields and summary statistics of the ensemble members. The output is the parameters of the postprocessed marginal distribution at each grid point (i.e., parameters of a GTCND or a CSGD). The architecture can be decomposed into two parts. On the left part, the succession of specific convolutional blocks (red and purple arrows) leads to an increase in the number of features and a reduction of the spatial dimension (i.e., a coarsening of the spatial resolution) as the data progresses through the network. As explained above, the convolutional blocks are constructed in order to learn useful representations of the features of the fields at various spatial scales. On the right part, upscaling blocks (red and orange arrows), based on bilinear upsampling, use the features learned in the central part of the architecture to predict features

at finer resolutions and finally learn the parameters of the distribution selected. Additionally, we use skip-connections (yellow arrows), consisting of copying and concatenating features, as bridges between the left and right parts of the U-Net. Skip-connections have proven to improve the stability of the convergence of NN (see, e.g., Li et al. 2018). This U-Net-based method is a global model enabling extrapolation through a parametric distribution (e.g., GTCND or CSGD). We denote U-Net+*distrib* the distributional regression U-Net (DRU) where *distrib* is the parametric distribution.

DRU learns to predict the parameters of a distribution by minimizing the CRPS at each grid point. Both the parameterized distribution and the scoring rule to minimize can be chosen to be suited to the variable of interest or to facilitate computations, thus making the architecture flexible. The convolution blocks allow the parameters of marginal distribution to be learned from neighboring grid points, potentially accounting for dependencies between grid points (Schefzik and Möller, 2018, Section 4.5). Moreover, the use of constant fields as input enables the convolutional layers to learn representations of these fields that are relevant to the postprocessing task at hand. This can be seen as a natural extension of the embedding module in DRN (Rasp and Lerch, 2018).

DRNs are built to bypass the limitations of the methods presented above. The model is global and uses the predictor fields of the whole grid, this construction enables the predicted marginals to be spatially consistent. Moreover, the use of convolutional layers facilitates the learning of relevant spatial features compared to DRN. Memory voracity is not an issue as the model is global and the number of parameters is contained. Finally, as highlighted previously, any parameterized distribution can be used as the output of DRU accounting for extrapolation and relevance to the target variable at hand. Table 3.3 summarizes the characteristics of the postprocessing methods studied in this article.

The U-Net-based method of this article is related to the one of Hu et al. (2023) in the sense that both approaches use U-Nets to predict the parameters of a distribution corresponding to the marginals of the variable of interest. The main differences between the approaches are the following: they studied daily precipitation accumulations, where we are interested in 3-h accumulated precipitation; they postprocess deterministic forecasts, where we postprocess ensemble forecasts; and finally, we use constant fields as additional predictors. Moreover, in terms of the number of years in the training data, our work (with only 3 years of training data) falls in a "gray area" where their U-Net-based method is outperformed by analog ensemble (Delle Monache et al., 2013), which is a simpler approach (Hu et al., 2023, Figure 11). Table 3.4 summarizes the characteristics of the different U-Net-based postprocessing methods available.

The following hyperparameters of the U-Net architecture have been selected using the training/validation dataset: the learning rate, the batch size and the number of epochs. The optimizer is Adam with default parameters (except for the learning rate) from its `Keras` implementation. In order to limit the number of parameters and prevent overfitting, the depth of the U-Net is kept at two levels (as shown in Fig. 3.3) and separable convolutions were used instead of standard ones. Moreover, in order to contain the variability due to random initialization, we aggregate forecast distributions of 10 models as recommended in Schulz and Lerch (2022a).

Most of the implementation was conducted in `Python` and the implementation of DRU is based on `Tensorflow` (Abadi et al., 2015) and `Keras` (Chollet et al., 2015). QRF and TQRF are implemented in `R` (R Core Team, 2023) using the `ranger` package (Wright and Ziegler, 2017).

	QRF	TQRF	DRN	DRU
Local/Global	local	local	global	global
Principles	grid point per grid point	grid point per grid point	embedding to learn from similar stations	constant fields and architecture aware of the gridded structure
Ability to extrapolate	✗	✓	✓	✓
Number of parameters	∼15.3 B	∼15.3 B	∼450,000	∼1,000,000
Storage necessary for prediction	splits of each tree and training data	splits of each tree and training data	parameters and architecture	parameters and architecture

Table 3.3: Comparison of the postprocessing methods mentioned in this study. The number of parameters is provided for hyperparameters selected by cross-validation on the training/validation data set and for the setup described in Section 3.2 (e.g., a 112×192 grid). In the case of DRN, an architecture similar to the one in Rasp and Lerch (2018) has been considered. *B* stands for billion.

3.4 Results

We provide a comparison of DRU to QRF, TQRF and the raw ensemble using verification tools targeting three different aspects of forecasts: verification of the overall performance with the CRPS, calibration and extreme events. First, we compare the performance of the postprocessing techniques in terms of their relative improvement compared to the raw ensemble and among themselves. This improvement is quantified in terms of continuous ranked probability skill score (CRPSS). Second, we assess the calibration of the postprocessed forecasts using rank histograms. Finally, the improvement of the postprocessing methods in terms of extreme forecasting is evaluated using receiver operating characteristic (ROC) curves for events corresponding to the exceedance of various thresholds.

3.4.1 Continuous ranked probability score

Since the postprocessing techniques considered act on the 1-dimensional marginals, the improvement and comparison of the postprocessing techniques can be done with univariate scoring rules. The continuous ranked probability score (CRPS; Matheson and Winkler 1976) is one of the most popular univariate scoring rules in weather forecasting and is defined as

$$\begin{aligned} \mathrm{CRPS}(F, y) &= \int_{\mathbb{R}} (F(z) - \mathbb{1}_{y \leq z})^2 \mathrm{d}z; && (3.1) \\ &= 2 \int_0^1 (\mathbb{1}_{y \leq F^{-1}(\alpha)} - \alpha)(F^{-1}(\alpha) - y) \mathrm{d}\alpha; && (3.2) \\ &= \mathbb{E}_F |X - y| - \frac{1}{2} \mathbb{E}_F |X - X'|, && (3.3) \end{aligned}$$

where the forecast F is assimilated to its cdf, F^{-1} is its quantile function and X and X' follow the distribution F. The CRPS is strictly proper on the set of measures with a finite first

	Grönquist et al. (2021)	Dai and Hemri (2021)	Horat and Lerch (2024)	Ben Bouallègue et al. (2024b)	Hu et al. (2023)	Pic et al. (2024)
Variable of interest	temperature, geopotential	cloud cover	temperature, 2-w precip.	temperature, 6-h precip.	24-h precip.	3-h precip.
Output	bias	samples from a cGAN	probability of classes	postprocessed ensemble members	parameters of a CSGD	parameters of a GTC-ND/CSGD
Lead times	48h	1-120h	2-4w	6-96h	0-4d	21h
Dataset: raw forecast	ECMWF-ENS10 *ensemble*	COSMO-E, ECMWF-IFS *ensemble*	ECMWF-IFS (S2S) *ensemble*	ECMWF-IFS *ensemble*	West-WRF *deterministic*	AROME-EPS *ensemble*
Dataset: obs.	ERA5	EUMETSAT	NOAA-CPC	ERA5	PRISM	ANTILOPE
Resolution	0.5°	0.02°	1.5°	1°	0.04°	0.025°
Training data range	17 years	3 years	20 years	19 years	2-30 years	3 years

Table 3.4: Comparison of the postprocessing methods relying on U-Nets.

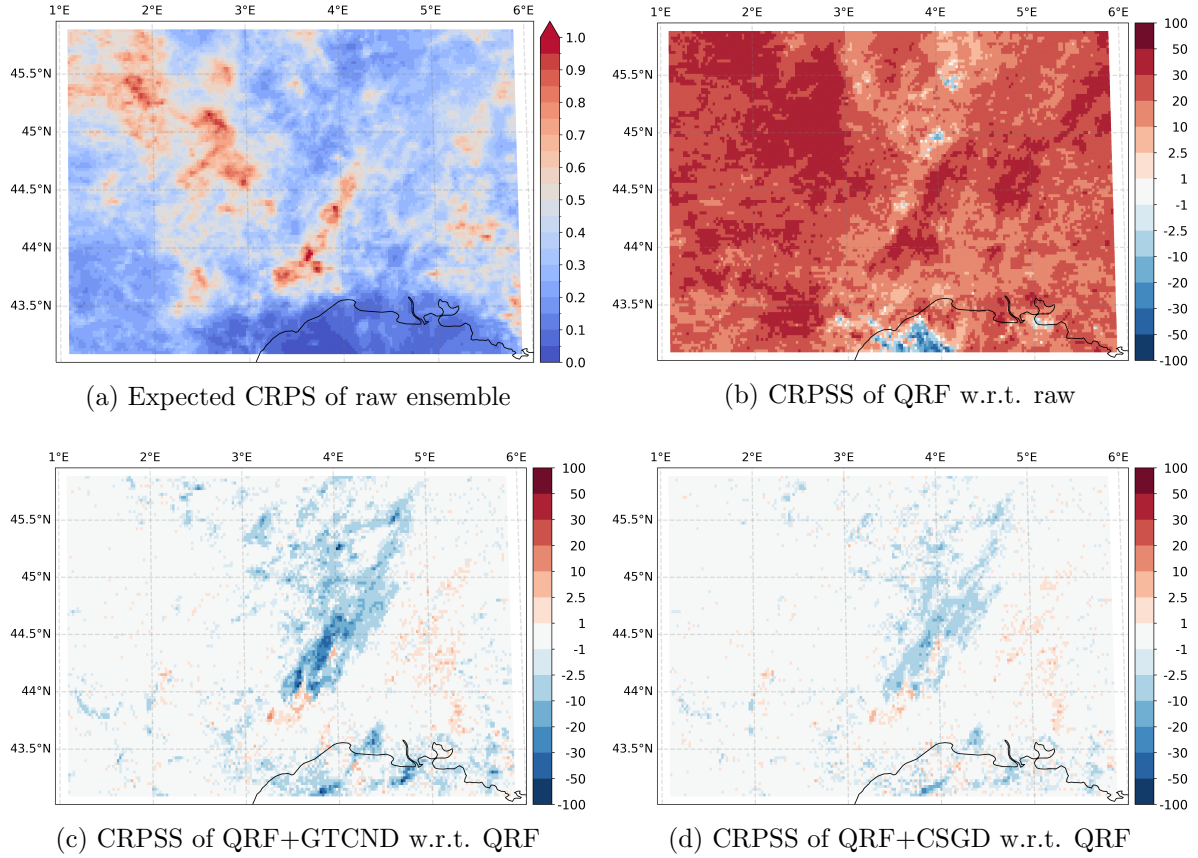
moment. Moreover, it benefits from multiple representations that help both its computation and interpretation. Equation (3.1) is the threshold or Brier score (Brier, 1950) representation and expresses the CRPS as the integrated squared error between the cdf of the forecast and the empirical cdf associated with observation y over all thresholds z. Equation (3.2) is the quantile representation and shows that the CRPS is expressed as the pinball loss over all quantile levels α. Equation (3.3) is the kernel representation and is particularly useful to compute the score of ensemble forecasts. The CRPS formulas for the parametric distributions of this article are available in the Appendix 3.6.1 and 3.6.2. For the raw ensemble, QRF and TQRF forecasts, the CRPS has been estimated using the fair estimator (Ferro, 2013).

When working with (strictly) proper scoring rules to compare forecasts, the comparison of the scoring rules of two forecasts can be summarized by the skill score. For a proper scoring rule S, the skill score of a forecast F with respect to (w.r.t.) a reference forecast F_{ref} is defined as

$$\text{SS}(F, F_{\text{ref}}) = \frac{\mathbb{E}_G[\text{S}(F_{\text{ref}}, Y)] - \mathbb{E}_G[\text{S}(F, Y)]}{\mathbb{E}_G[\text{S}(F_{\text{ref}}, Y)]}, \tag{3.4}$$

where G is the distribution of the observations and $\mathbb{E}_G[\cdots]$ is the expectation with respect to $Y \sim G$. The skill score is positive if the forecast F improves the expected score w.r.t. the reference forecast F_{ref} and negative otherwise. The skill score can be expressed in percentage. In the context of postprocessing, a reference of choice is the raw ensemble that the postprocessing procedure aims to improve upon.

We compared the continuous ranked probability skill score (CRPSS) for the different postprocessing methods studied w.r.t. other benchmark methods. Figure 3.4 shows the expected CRPS of the raw ensemble, the CRPSS of QRF w.r.t. the raw ensemble and the CRPSS of QRF+GTCND and QRF+CSGD w.r.t. QRF. The raw ensemble has an expected CRPS of 0.3725 mm when averaged over the whole region of interest. However, the expected CRPS greatly fluctuates over the whole grid and most grid points of higher altitude have larger expected CRPS since they correspond to higher precipitation accumulations (see Fig. 3.4a). The



(a) Expected CRPS of raw ensemble

(b) CRPSS of QRF w.r.t. raw

(c) CRPSS of QRF+GTCND w.r.t. QRF

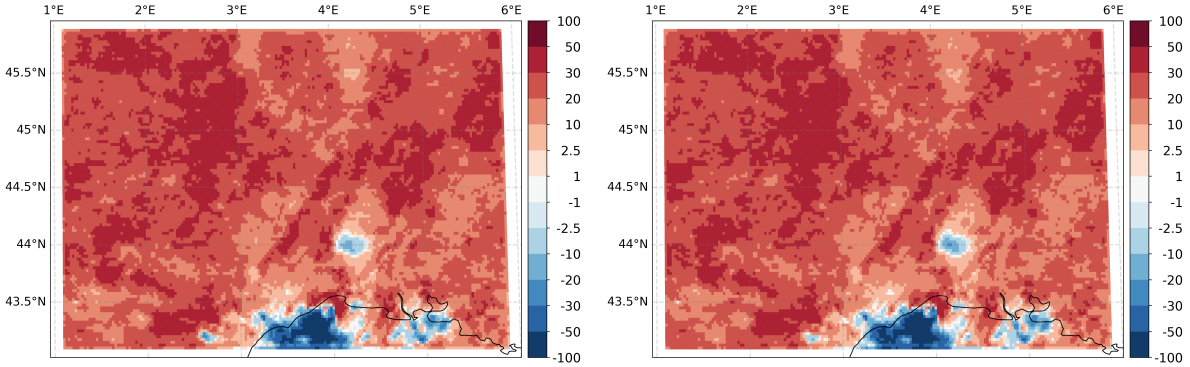
(d) CRPSS of QRF+CSGD w.r.t. QRF

Figure 3.4: Predictive performance of the benchmark methods in terms of CRPS. (a) Expected CRPS of the raw ensemble, (b) CRPSS of QRF w.r.t. the raw ensemble and CRPSS w.r.t. QRF of (c) QRF+GTCND and (d) QRF+CSGD.

lowest expected CRPS values are located over the Mediterranean Sea corresponding to an area of low precipitation as discussed further (see Fig. 3.6). Moreover, observations in this area are of lower quality since it is far from the nearest radar and cannot be corrected by gauges.

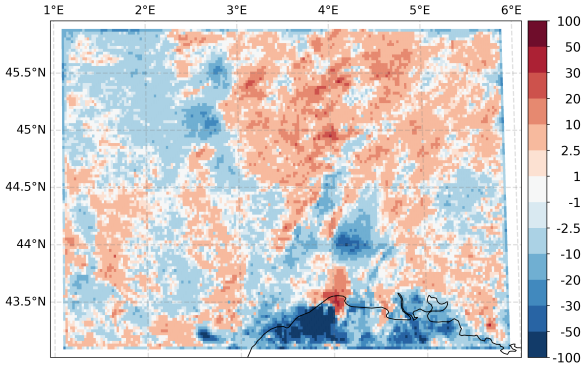
Figure 3.4b confirms that QRF is able to improve the predictive performance in terms of CRPSS compared to the raw ensemble (23.51% after averaging over the region of interest). The CRPSS of QRF w.r.t. the raw ensemble is positive (i.e., improvement of skill) over the whole domain except for some localized regions. In particular, over the area that has the lowest expected CRPS for the raw ensemble, QRF is not able to improve compared to raw ensemble in terms of expected CRPS. This may be caused by the fact that this area is already well-predicted by the raw ensemble and the QRF is not able to improve its CRPS. Figures 3.4c and 3.4d show the CRPSS w.r.t. QRF of QRF+GTCND and QRF+CSGD, respectively. Overall, QRF+GTCND and QRF+CSGD have a close but slightly smaller expected CRPS than that of QRF (average CRPSS w.r.t. QRF of -1.04% and -0.33%, respectively). For both GTCND and CSGD tail extensions, the areas of lower skill (in blue) are located in a mountainous region (the Eastern part of Massif Central) and near the Mediterranean coast. Nonetheless, the areas are wider and have lower CRPSS values for QRF+GTCND compared to QRF+CSGD. Both methods also present areas of improvement of CRPSS (in orange/red) that are sparser and smaller than the areas of negative CRPSS.

Figure 3.5 provides the CRPSS of U-Net+GTCND and U-Net+CSGD w.r.t. the raw ensemble and QRF. Figures 3.5a and 3.5b show the CRPSS of DRU w.r.t. the raw ensemble. Both GTCND and CSGD lead to methods improving CRPSS w.r.t. the raw ensemble with

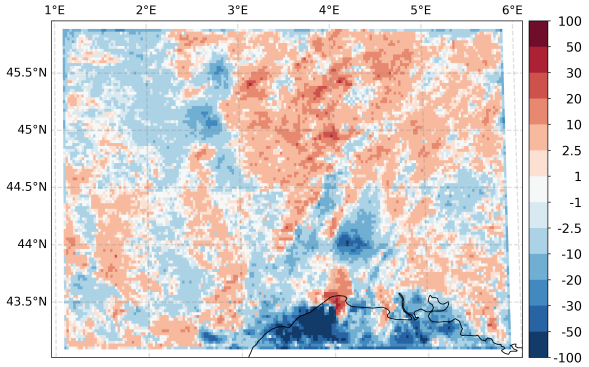


(a) CRPSS of U-Net+GTCND w.r.t. raw

(b) CRPSS of U-Net+CSGD w.r.t. raw



(c) CRPSS of U-Net+GTCND w.r.t. QRF



(d) CRPSS of U-Net+CSGD w.r.t. QRF

Figure 3.5: Predictive performance of the distributional regression U-Nets in terms of CRPS. CRPSS w.r.t. the raw ensemble of (a) U-Net+GTCND and (b) U-Net+CSGD and CRPSS w.r.t. QRF of (c) U-Net+GTCND and (d) U-Net+CSGD.

22.28% and 22.36%, respectively, when averaged over the region of interest. As the QRF, DRU leads to improvement in terms of CRPSS over the vast majority of grid points. Nonetheless, there are areas where they have a poorer predictive performance compared to raw ensemble. These areas are also located over the Mediterranean Sea or near the coast, and one patch is located in the Rhône River valley. When censoring grid points located over the sea and at the border, the average CRPSS w.r.t. the raw ensemble is 24.34% and 24.48% for U-Net+GTCND and U-Net+CSGD, respectively.

Figures 3.5c and 3.5d show the CRPSS of U-Net+GTCND and U-Net+CSGD w.r.t. QRF. Overall, DRU has a higher expected CRPS than QRF (CRPSS of -1.52% for the U-Net+GTCND and -1.37% for the U-Net+CSGD), but it has an improved predictive performance (in terms of CRPS) over a non-negligible part of the region of interest. Due to their architecture, DRUs are affected by a border effect, leading to a less predictive performance on the grid points located at the boundaries of the grid (see Fig. 3.5c and Fig. 3.5d). Using the censoring mentioned above, U-Net+GTCND and U-Net+CSGD have an average CRPSS w.r.t. QRF of 0.05% and 0.26%, respectively. Table 3.5 summarizes the comparisons of methods in terms of CRPSS.

For the training/validation dataset, DRUs are prone to numerical instabilities. This led to areas of negative CRPSS w.r.t. the raw ensemble caused by the divergence of predicted parameters (σ is the case of U-Net+GTCND and θ in the case of U-Net+CSGD) (not shown). In addition to standard numerical stabilizing tricks, we have tried to constrain the range of diverging parameters using the value of the climatological fits since higher values would lead to forecasts less informative than the climatological forecasts. This solved the divergence issues over both

		Reference			
		Full region		Censored region	
		Raw ensemble	QRF	Raw ensemble	QRF
Postprocessing methods	QRF	**23.51%**	–	23.56%	–
	QRF+GTCND	22.67%	-1.04%	22.72%	-1.05%
	QRF+CSGD	23.23%	-0.33%	23.29%	-0.34%
	U-Net+GTCND	22.25%	-1.52%	24.34%	0.05%
	U-Net+CSGD	22.36%	-1.37%	**24.48%**	**0.26%**

Table 3.5: Summary of the performance in terms of CRPSS averaged over the full region of interest and over the censored one.

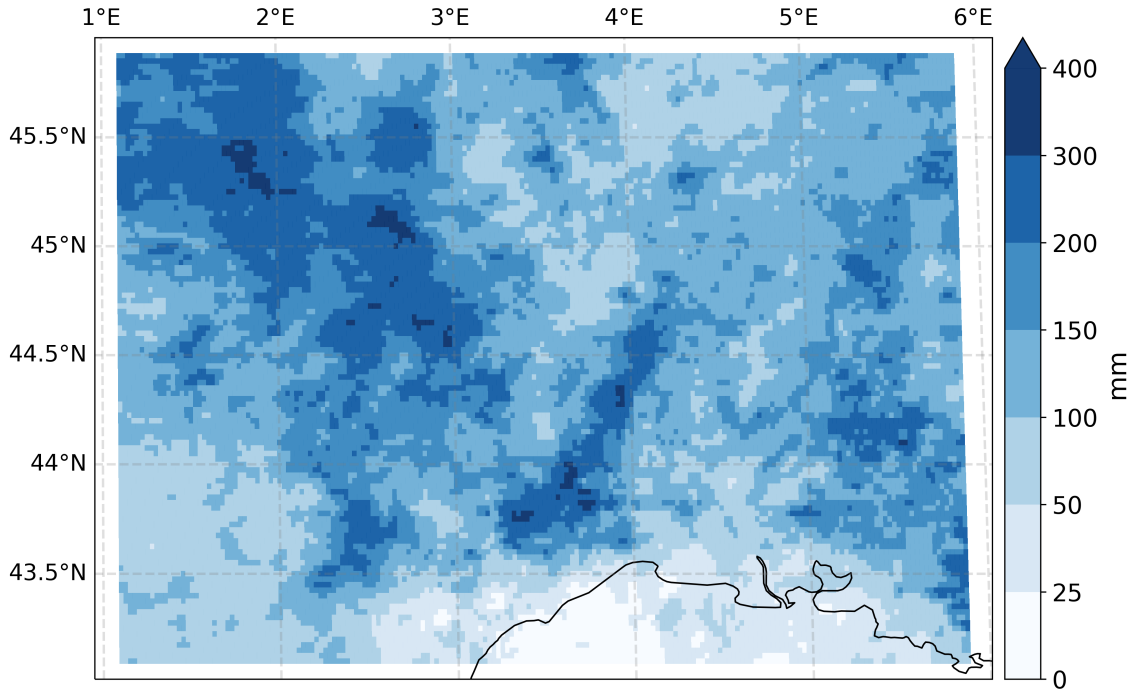
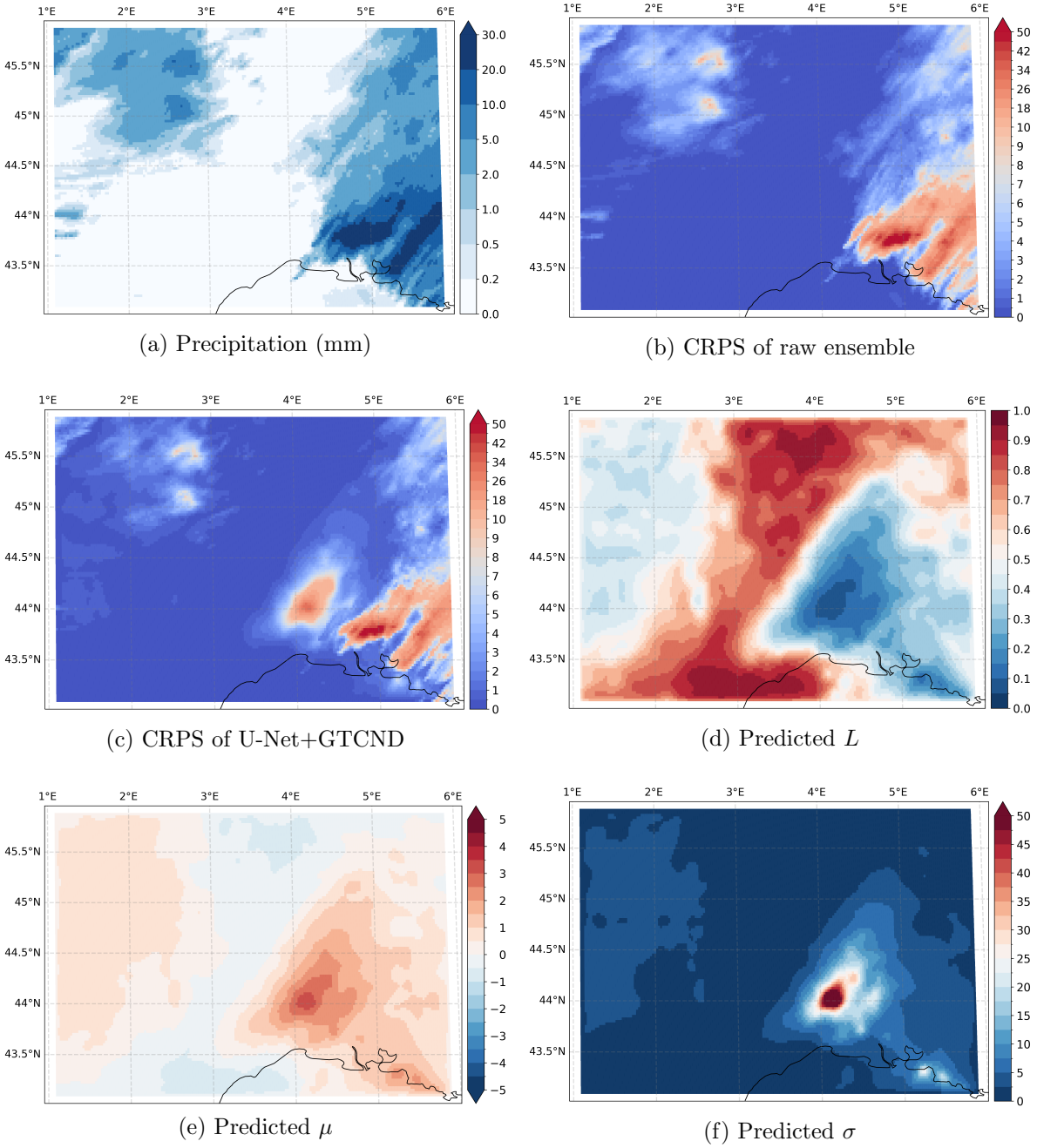


Figure 3.6: Total precipitation over the test set. Due to the initial time and the lead time considered, only precipitation between 12:00UTC and 15:00UTC are taken into account.

the training/validation and test datasets for U-Net+CSGD but not for U-Net+GTCND (not shown). However, it increased the border effects causing deteriorating performance for both models. Hence, the constraining of the range of the parameters for DRU method is not used and the numerical stability of the methods needs to be understood and prevented.

Despite being prone to numerical instabilities, the areas of negative CRPSS w.r.t. the raw ensemble for the test dataset are not all caused by numerical instabilities. The largest area of negative CRPSS w.r.t. raw (see Fig. 3.5a and 3.5b) coincide with the area with the lowest total precipitation over the test period (see Fig. 3.6). This area matches the area of the lowest expected CRPS for the raw ensemble (see Fig. 3.4a). Numerous dry events occur at this location and are perfectly predicted by the raw ensemble (i.e., all members predict 0 mm of precipitation).



(a) Precipitation (mm)

(b) CRPS of raw ensemble

(c) CRPS of U-Net+GTCND

(d) Predicted L

(e) Predicted μ

(f) Predicted σ

Figure 3.7: Example of a numerical instability of U-Net+GTCND for a forecast valid on November 3, 2022 at 15 :00UTC. Note the different scales for CRPS below and above 10 mm.

However, in order to perfectly predict a dry event, U-Net+GTCND and U-Net+CSGD need to predict $L = 1$ and $-\delta/\theta = \infty$, respectively, which is never the case in practice. This may explain why the CRPSS w.r.t. the raw ensemble of this area is highly negative for DRU. The CRPS of QRF (and TQRF) has been computed using 107 quantiles, rendering perfect prediction of dry events harder and resulting in a deterioration in terms of CRPS over the aforementioned area (see Fig. 3.4b).

The other smaller areas of negative CRPSS w.r.t. the raw ensemble for DRU seem to be caused by numerical instabilities. For example, Figure 3.7 presents a numerical instability for a U-Net+GTCND forecast valid on November 3, 2022 at 12:00UTC. It corresponds to heavy precipitation over the Easter part of the region of interest (see Fig. 3.7a). Both raw ensemble and U-Net+GTCND seem not able to correctly predict heavy precipitation, as reflected in the

high values of their CRPS (see Fig. 3.7b and 3.7c). However, the CRPS of U-Net+GTCND presents an additional area of high CRPS that is caused by the prediction of precipitation where no precipitation has been observed. This incorrect prediction is characterized by a low value of L (i.e., low probability of dry event), a positive value of μ and a very high value of σ (see Fig. 3.7d, 3.7e and 3.7f). The abnormally large value of σ seems to be caused by a numerical instability and gives a larger probability to large precipitation. The high CRPS over this region associated with a low value of CRPS for raw ensemble causes the CRPSS for U-Net+GTCND w.r.t. the raw ensemble over the test set to be negative (see Fig. 3.5a).

DRUs are able to reach a predictive performance slightly lower but comparable to the QRF. U-Net+CSGD has a slightly better expected CRPS than U-Net+GTCND. In order to be deemed worthy postprocessing methods, U-Net+GTCND and U-Net+CSGD need to be calibrated.

3.4.2 Calibration

Since the ideal forecast (i.e., the true conditional distribution) is unknown, it is impossible to know if a postprocessed forecast has reached the minimum expected CRPS. In order to decompose the contribution of calibration and sharpness to scoring rules (Winkler, 1977; Winkler et al., 1996), rank histograms are used to evaluate the calibration of the different postprocessing techniques.

Multiple definitions of calibration exist with different levels of hypotheses (see, e.g., Tsyplakov 2013, 2020). The most used definition is probabilistic calibration which, broadly speaking, consists of computing the rank of observations among samples of the forecast and checking for uniformity with respect to observations. If the forecast is calibrated, observations should not be distinguishable from forecast samples, and thus, the distribution of their ranks should be uniform, leading to a flat histogram. The shape of the rank histogram gives information about the type of (potential) miscalibration: a triangular-shaped histogram suggests that the probabilistic forecast has a systematic bias, a $\cup$-shaped histogram suggests that the probabilistic forecast is underdispersed and a $\cap$-shaped histogram suggests that the probabilistic forecast is overdispersed. Jolliffe and Primo (2008) proposed a statistical test to assess the uniformity (i.e., flatness) of rank histograms. Moreover, slopes in the rank histograms can be accounted for. Zamo (2016) proposed a test accounting for the presence of a wave in rank histograms. This test is called the Jolliffe-Primo-Zamo (JPZ) test in the following.

To conciliate with the AROME-EPS raw ensemble composed of 17 members, the rank histograms can take 18 different classes and 107 quantiles of the forecasts were produced for the QRF, TQRF and DRU methods (each group of 6 consecutive ranks are gathered as a single rank).

Figure 3.8 shows the rank histograms of each forecast over the whole grid and the JPZ tests for flatness of rank histograms. As is often the case, the raw ensemble is biased and underdispersed, which is visible by the triangular shape of the rank histograms and the fact that the lowest and highest ranks are over-represented. Its JPZ test confirms that the raw ensemble forecast is not calibrated (only 6% of grid points do not reject the flatness of the rank histogram). QRF, QRF+GTCND and QRF+CSGD all show very high calibration with JPZ tests not rejecting flatness at 93%, 94% and 93% of grid points. Contrary to what was observed in Taillardat et al. (2019), no noticeable difference in calibration seems to be present between the QRF and its tail extension. This may be caused by the operational refinement used in the implementation, the fact that different parametric distributions are used and the smaller precipitation accumulations compared to the original article (i.e., 3-h vs. 6-h). DRUs present a lower calibration level compared to QRF-based methods, but their calibration is still significant. The JPZ tests do not reject the flatness hypothesis at 74% and 77% of the grid points for the U-Net+GTCND and U-Net+CSGD, respectively. Both DRU forecasts present a slight underdispersion in the right tail revealed by the higher representation of the largest rank

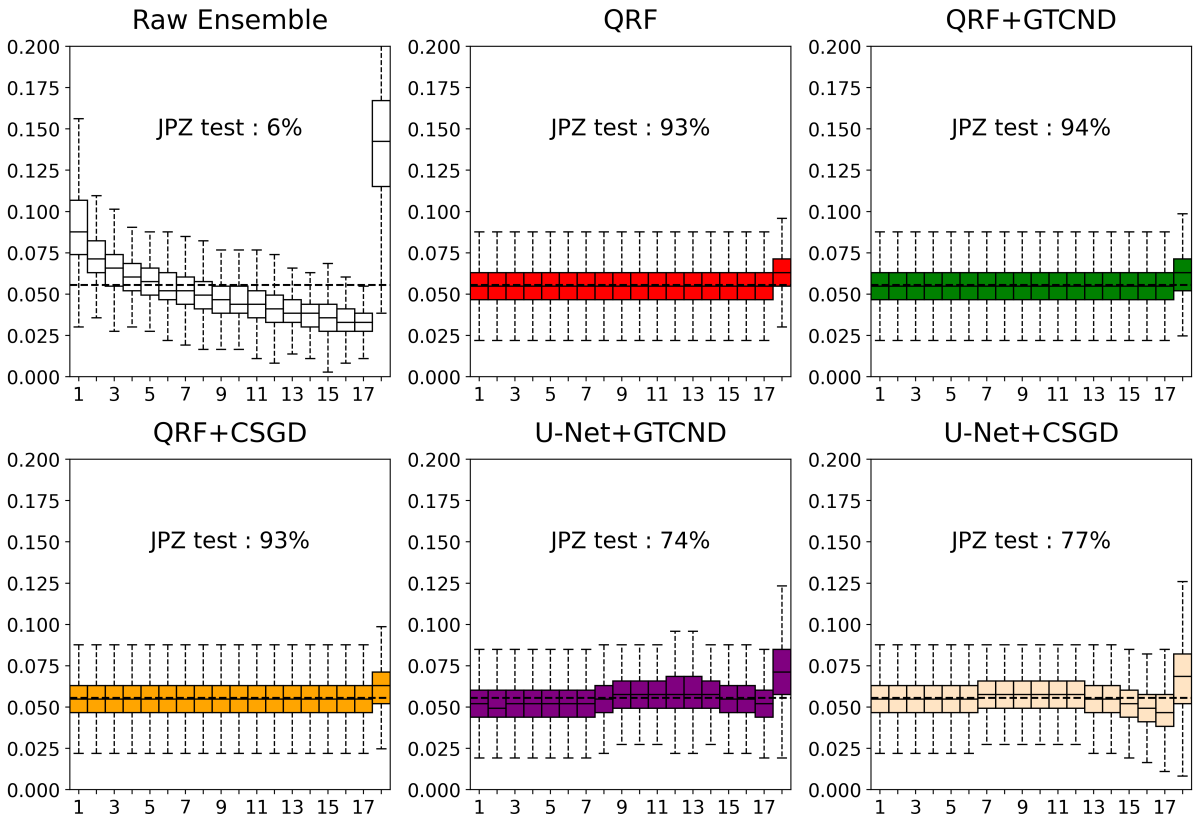


Figure 3.8: Rank histogram for raw ensemble, QRF, TQRF (namely, QRF+GTCND and QRF+CSGD) and distributional regression U-Nets associated with the GTCND and the CSGD. The hyperparameters are selected as the best performing by cross-validation on the training dataset.

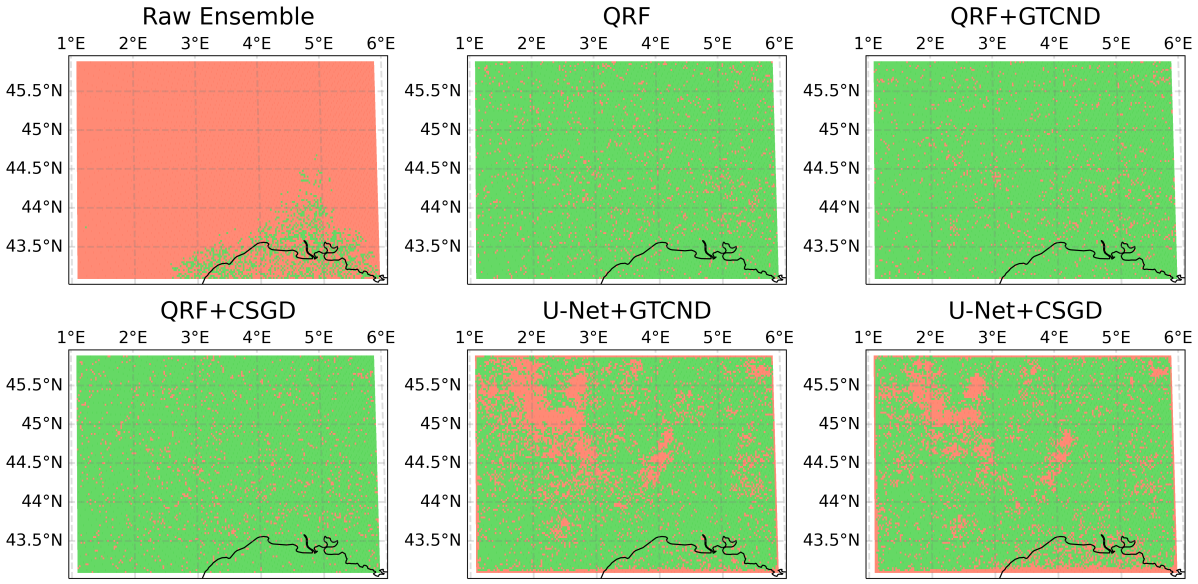


Figure 3.9: Map of rejection (red) and non-rejection (green) of the flatness of the rank histogram for the forecasting methods considered: raw ensemble, QRF, QRF+GTCND, QRF+CSGD, U-Net+GTCND and U-Net+CSGD.

in their histograms.

Figure 3.9 shows a map of the rejection and non-rejection of the flatness of the rank histogram given by JPZ tests. Calibrated grid points for the raw ensemble are sparsely located over the Mediterranean Sea, the coast and the South of the Rhône valley. QRF, QRF+GTCND and QRF+CSGD are able to calibrate the marginals homogeneously across the region of interest. The areas explaining the lower rate of calibrated grid point for DRU compared to QRF-based methods correspond to high climatological precipitation (see Fig. 3.6). The lack of calibration over these areas may be caused by the small depth of the training/validation data (only 3 years) resulting in not enough high precipitation observed. Moreover, the lower performance due to border effects affects the calibration of the DRU forecasts. DRU leads to spatially inconsistent forecasts in terms of calibration whereas the QRF-based methods are homogeneously calibrated over the whole domain.

3.4.3 Extreme events

Extreme events are of particular interest. They may lead to the highest socio-economic impacts. However, if verification were to focus only on cases of extreme events, forecasters might be encouraged to propose forecasts that are overly alarming and, thus, of lower general predictive performance. Lerch et al. (2017) pinpointed this phenomenon and named it the *forecaster's dilemma*. Since we have compared the general predictive performance of postprocessing techniques, we can conduct verification focused on extreme events and not be affected by the forecaster's dilemma.

To focus on forecasts' predictive performance regarding extreme events, we are interested in predicting binary events in the form of the exceedance of a high threshold t. We use ROC (receiver operating characteristic) curves to evaluate the discriminant power of forecasts in terms of binary decisions. In particular, ROC curves can inform on the risk of missing an extreme event. Given the binary event $\mathbb{1}_{y>t}$ (i.e., exceedance of the threshold t), the ROC curve is the plot of the rate of predicted events (i.e., true positive), also called hit rate, versus the rate of false alarms (i.e., false positive). A good forecast should maximize the rate of events detected while minimizing false alarms. In practice, the compromise between the highest hit rate of the method and its lowest false alarm rate depends on the application. In the case of high-impact events, forecasts with a non-negligible false alarm rate may be tolerated if it is accompanied by a better hit rate. In addition to thresholds associated with extreme events, lower thresholds corresponding to lower precipitation events are investigated. Note that grid point by grid point computation of ROC curves does not prevent potential double-penalty effects (Ebert, 2008).

In Figure 3.10, ROC curves for the exceedance of various thresholds are represented for the raw ensemble, QRF, TQRF and DRU. The lowest threshold is $t = 0$mm, which characterizes the prediction of dry events (i.e., absence of precipitation). Raw ensemble has a poor performance regarding the prediction of the presence of precipitation. All the postprocessing methods have comparable performances, as seen in the overlap of their ROC curves. During the cross-validation over the training/validation dataset, the raw ensemble had a better predictive power regarding the prediction of dry events but was still lower than the postprocessing methods (not shown). The threshold $t = 5$ mm corresponds to intermediate precipitations. The performance of the raw ensemble already decreases and a difference between DRU and QRF-based methods appears. DRUs have a slightly higher predictive performance compared to QRF-based methods. The raw ensemble lacks resolution because of the nature of its miscalibration (i.e., bias and underdispersion).

For the highest thresholds $t = 10$ mm and $t = 20$ mm (corresponding to the quantile of level 0.995 and 0.999, respectively, of the climatology over the region of interest), the ROC curves of the different postprocessing methods can be distinguished. For $t = 10$ mm, the performance of the raw ensemble continues to deteriorate and is close to the random guess (dashed line).

All the postprocessing techniques are able to maintain a good predictive power but start to noticeably lack resolution, which can be seen in the sudden change of slope. U-Net+GTCND and U-Net+CSGD have a better performance compared to QRF-based techniques which continue to have overlapping ROC curves. U-Net+CSGD has the overall best performance. For $t = 20$ mm, the raw ensemble has a performance indistinguishable from a random guess. DRUs are better than QRF-based methods. QRF+GTCND and QRF+CSGD denote from QRF as the tail extension improves predictive performance. QRF+GTCND seems to have a slightly better performance than QRF+CSGD. The gap in performance between U-Net+CSGD and U-Net+GTCND continues to grow and U-Net+CSGD clearly has to the best predictive power w.r.t. the exceedance of the threshold $t = 20$ mm.

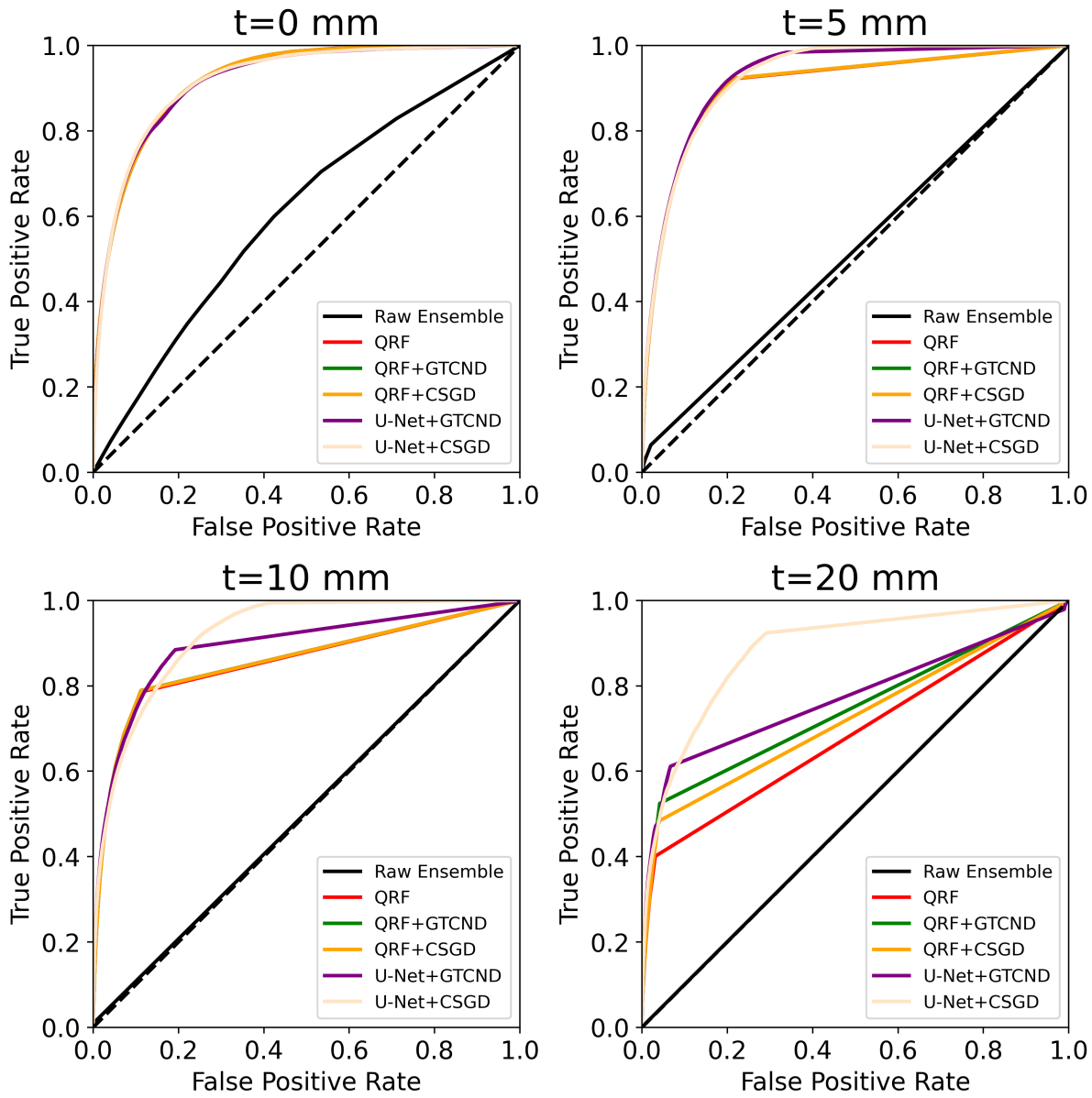


Figure 3.10: Receiver operating characteristic (ROC) curves of binary events corresponding to the exceedance of a threshold $t \in \{0, 5, 10, 20\}$ (in mm of precipitation). As for Figure 3.8, the hyperparameters are selected as the best performing by cross-validation on the training dataset.

All postprocessing methods compete favorably with the raw ensemble, which has the same

predictive performance as a random guess for the highest thresholds ($t = 10$ mm and $t = 20$ mm). All postprocessing methods have comparable predictive performances for dry events. For heavy precipitation events corresponding to quantiles of levels 0.995 and 0.999, DRUs, and in particular U-Net+CSGD, have a distinctly better predictive power. Moreover, as already observed in Taillardat et al. (2016), TQRF is able to improve the prediction of heavy precipitation with respect to QRF (even for a light-tailed extension as the GTCND).

3.5 Discussion

We proposed a U-Net-based method, namely distributional regression U-Nets, to postprocess marginal distributions for gridded precipitation data. This approach extends DRN to gridded data by substituting the fully connected NN and embedding module for a U-Net architecture aware of the gridded structure of the data. Simultaneously predicting marginal distributions at each grid point using information from nearby grid points represents a means to account for dependencies between grid points. Both U-Net+GTCND and U-Net+CSGD have predictive performances comparable to the QRF and TQRF in terms of CRPS. DRUs are (probabilistically) calibrated over a large part of the domain studied except for areas associated with the highest precipitation over the test set (see Fig. 3.6). This may result from the relatively small training/validation set and could improve with a larger training/validation set. Future studies could try to limit this by emphasizing the learning of high precipitation events using weighted scoring rules for inference. In terms of heavy precipitation, U-Net+CSGD outperforms QRF-based methods.

One of the challenges of the dataset used is the small amount of available training data. This is encountered in practice where consistent data is required, but large reforecast and reanalysis are too computationally expensive. In a more general context, the lack of consistency can be induced at larger time scales by climate change or in specific regions of the world by El Niño forcing.

We focused on distributional regression U-Nets where outputs are distribution parameters based on CRPS minimization. DRU can rely on the minimization of other (strictly) proper scoring rules. Moreover, DRU can directly be extended to learn nonparametric distributions such as BQN (Bremnes, 2020) where the quantile function is a combination of Bernstein polynomials or as HEN (e.g., Scheuerer et al. 2020) where the pdf is modeled by the probability of bins.

As U-Net architecture is aware of the spatial gridded structure of the data, specific architectures can also be used for common data structures. We present architectures related to temporal and graph-based structures that are currently used in probabilistic forecasting settings. Their application to postprocessing provides an interesting for future works. For example, if the temporal structure of the data is of interest, recurrent neural networks can be used to predict a parametric distribution. Pasche and Engelke (2024) proposed to forecast flood risk using high-quantile prediction based on fitting a generalized Pareto distribution via logarithmic score (i.e., negative log-likelihood) minimization. In the case of spatial structure relying on an irregular or more abstract grid (e.g., station network), graph neural networks (GNNs) are able to predict graph-based quantities (Battaglia et al., 2018). Cisneros et al. (2024) used graph convolutional neural networks to learn the parameters of a mixture of a logistic distribution and EGPD via logarithm score minimization to predict wildfire spread. Using the 3D spatial graph-based structures, GNNs are already able to produce deterministic forecasts reaching performance comparable to ECMWF deterministic high-resolution forecasts in performance (Keisler, 2022; Pathak et al., 2022; Bi et al., 2023; Lam et al., 2023; Chen et al., 2023).

Acknowledgments

The authors acknowledge the support of the French Agence Nationale de la Recherche (ANR) under reference ANR-20-CE40-0025-01 (T-REX project) and the Energy-oriented Centre of Excellence II (EoCoE-II), Grant Agreement 824158, funded within the Horizon2020 framework of the European Union. Part of this work was also supported by the ExtremesLearning grant from 80 PRIME CNRS-INSU and this study has received funding from Agence Nationale de la Recherche - France 2030 as part of the PEPR TRACCS program under grant number ANR-22-EXTR-0005 and the ANR EXSTA.

3.6 Appendix

3.6.1 Generalized Truncated/Censored Normal Distribution

We recall quantities related to the generalized truncated/censored normal distribution (GTCND). Denote l and u the lower and upper boundaries, L and U are the point masses at these boundaries. Since we are working with precipitation, we are interested in the case where $u = \infty$ (implying that $U = 0$) and $l = 0$, leaving L a parameter to determine along μ and σ. Formulas for the general case are available in Jordan et al. (2019).

The cumulative distribution function (cdf) of the GTCND is

$$F_{L,\mu,\sigma}^{gtcnd}(z) = \begin{cases} \dfrac{1-L}{1-\Phi(-\mu/\sigma)}\big(\Phi(\frac{z-\mu}{\sigma}) - \Phi(-\mu/\sigma)\big) + L & \text{if } z \geq 0 \\ 0 & \text{if } z < 0 \end{cases}$$

where Φ is the cdf of the standard normal distribution. Its quantile function is expressed as

$$F_{L,\mu,\sigma}^{\text{gtcnd}^{-1}}(p) = \begin{cases} 0 & \text{if } p \leq L \\ \mu + \sigma\Phi^{-1}\left(\frac{(p-L)(1-\Phi(-\mu/\sigma)}{1-L} + \Phi(-\mu/\sigma)\right) & \text{if } p > L \end{cases}$$

for $p \in (0, 1)$. The special case of GTCND used here can be expressed using the truncated normal distribution :

$$F_{L,\mu,\sigma}^{\text{gtcnd}}(z) = L\mathbb{1}_{z\geq 0} + (1-L)N_{\mu,\sigma}^{0}(z),$$

where $N_{\mu,\sigma}^{0}$ is the cdf of the zero-truncated normal distribution.

Moments methods

$$\mathbb{E}[\mathbb{1}_{X=0}] = L$$

$$\mathbb{E}[X] = \mu + \frac{\phi(-\mu/\sigma)\sigma}{1-\Phi(-\mu/\sigma)}$$

$$\mathbb{V}\text{ar}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \sigma^2\left\{1 - \frac{\mu}{\sigma}\frac{\phi(\mu/\sigma)}{1-\Phi(-\mu/\sigma)} - \left(\frac{\phi(-\mu/\sigma)}{1-\Phi(-\mu/\sigma)}\right)^2\right\}$$

Continuous Ranked Probability Score

$$\begin{aligned}\mathrm{CRPS}(F^{\mathrm{gtcnd}}_{L,\mu,\sigma}, y) = |y - y_+| + \mu L^2 \\ + \frac{1-L}{1-\Phi(-\frac{\mu}{\sigma})}(y_+ - \mu)\left\{2\Phi\left(\frac{y_+-\mu}{\sigma}\right) - \frac{1-2L+\Phi\left(-\frac{\mu}{\sigma}\right)}{1-L}\right\} \\ + 2\sigma\frac{1-L}{1-\Phi(-\frac{\mu}{\sigma})}\left(\phi\left(\frac{y_+-\mu}{\sigma}\right) - \phi\left(-\frac{\mu}{\sigma}\right)L\right) \\ - \left(\frac{1-L}{1-\Phi(-\frac{\mu}{\sigma})}\right)^2 \frac{\sigma}{\sqrt{\pi}}\Phi\left(\frac{\mu\sqrt{2}}{\sigma}\right)\end{aligned}$$

with $y_+ = \max(0, y)$ and ϕ the probability density function of the standard normal distribution.

3.6.2 Censored-Shifted Gamma Distribution

We recall quantities related to the censored-shifted gamma distribution (CSGD). The expressions can be found in Scheuerer and Hamill (2015a) and Baran and Nemoda (2016). The cumulative distribution function (cdf) of the CSGD is

$$F^{\mathrm{csgd}}_{k,\theta,\delta}(z) = \begin{cases} G_k(\frac{z-\delta}{\theta}) & \text{if } z \geq 0 \\ 0 & \text{if } z < 0 \end{cases},$$

with G_k the cdf of the gamma distribution of shape k. Its quantile function is expressed as

$$F^{\mathrm{csgd}^{-1}}_{k,\theta,\delta}(p) = \delta + \theta\gamma^{-1}(k, p\Gamma(k)),$$

where γ is the lower incomplete gamma function, Γ is the gamma function and $p \in (0, 1)$.

Moments method

Let $\tilde{c} = -\delta/\theta$.

$$\mathbb{E}[X] = (1 - G_k(\tilde{c}))\Big\{\theta k(1 - G_{k+1}(\tilde{c})) - \delta(1 - G_k((\tilde{c}))\Big\}$$

$$\begin{aligned}\mathbb{E}[X^2] = (1 - G_k((\tilde{c}))\Big\{&k(k+1)\theta^2(1 - G_{k+2}(\tilde{c})) \\ &- 2\delta k\theta(1 - G_{k+1}(\tilde{c})) \\ &+ \delta^2(1 - G_k(\tilde{c}))\Big\}\end{aligned}$$

$$\begin{aligned}\mathbb{E}[X^3] = (1 - G_k(\tilde{c})\Big\{&k(k+1)(k+2)\theta^3(1 - G_{k+3}(\tilde{c})) \\ &- 3\delta k(k+1)\theta^2(1 - G_{k+2}(\tilde{c})) \\ &+ 3\delta^2 k\theta(1 - G_{k+1}(\tilde{c})) \\ &- \delta^3(1 - G_k(\tilde{c}))\Big\}\end{aligned}$$

Continuous Ranked Probability Score

The continuous ranked probability score (CRPS) of the CSGD is

$$\begin{aligned}\mathrm{CRPS}(F^{\mathrm{csgd}}_{k,\theta,\delta}, y) = \theta\Big\{&\tilde{y}\left(2G_k(\tilde{y}) - 1\right) - \tilde{c}G_k^2(\tilde{c}) + \theta k\left(1 + 2G_k(\tilde{c})G_{k+1}(\tilde{c}) - G_k^2(\tilde{c}) - 2G_{k+1}(\tilde{y})\right) \\ &- \frac{\theta k}{\pi}B(1/2, k+1/2)\left(1 - G_{2k}(2\tilde{c})\right)\Big\},\end{aligned}$$

where $\tilde{y} = \frac{y-\delta}{\theta}$, $\tilde{c} = -\delta/\theta$ and B is the beta function.

Chapter 4

Proper scoring rules for multivariate probabilistic forecasts based on aggregation and transformation

This chapter reproduces an article submitted to *Advances in Statistical Climatology, Meteorology and Oceanography*, and written by Romain Pic[1], Clément Dombry[1], Philippe Naveau[2] and Maxime Taillardat[3].

Abstract Proper scoring rules are an essential tool to assess the predictive performance of probabilistic forecasts. However, propriety alone does not ensure an informative characterization of predictive performance and it is recommended to compare forecasts using multiple scoring rules. With that in mind, interpretable scoring rules providing complementary information are necessary. We formalize a framework based on aggregation and transformation to build interpretable multivariate proper scoring rules. Aggregation-and-transformation-based scoring rules are able to target specific features of the probabilistic forecasts; which improves the characterization of the predictive performance. This framework is illustrated through examples taken from the literature and studied using numerical experiments showcasing its benefits. In particular, it is shown that it can help bridge the gap between proper scoring rules and spatial verification tools.

Contents

[1]Université de Franche Comté, CNRS, LmB (UMR 6623), F-25000 Besançon, France
[2]Laboratoire des Sciences du Climat et de l'Environnement, UMR 8212, CEA-CNRS-UVSQ, EstimR, IPSL & U Paris-Saclay, Gif-sur-Yvette, France
[3]CNRM, Université de Toulouse, Météo-France, CNRS, Toulouse, France

4.1 Introduction

Probabilistic forecasting allows to issue forecasts carrying information about the prediction uncertainty. It has become an essential tool in numerous applied fields such as weather and climate prediction (Vannitsem et al., 2021; Palmer, 2012), earthquake forecasting (Jordan et al., 2011; Schorlemmer et al., 2018), electricity price forecasting (Nowotarski and Weron, 2018) or renewable energies (Pinson, 2013; Gneiting et al., 2023) among others. Moreover, it is slowly reaching fields further from "usual" forecasting, such as epidemiology predictions (Bosse et al., 2023) or breast cancer recurrence prediction (Al Masry et al., 2023). In weather forecasting, probabilistic forecasts often take the form of ensemble forecasts in which the dispersion among members captures forecast uncertainty.

The development of probabilistic forecasts has induced the need for appropriate verification methods. Forecast verification fulfills two main purposes: quantifying how good a forecast is given observations available and allowing one to rank different forecasts according to their predictive performance. Scoring rules provide a single value to compare forecasts with observations. Propriety is a property of scoring rules that encourages forecasters to follow their true beliefs and that prevents hedging. Proper scoring rules allow to assess calibration and sharpness simultaneously (Winkler, 1977; Winkler et al., 1996). Calibration is the statistical compatibility between forecasts and observations. Sharpness is the uncertainty of the forecast itself. Propriety is a necessary property of good scoring rules, but it does not guarantee that a scoring rule provides an informative characterization of predictive performance. In univariate and multivariate settings, numerous studies have proven that no scoring rule has it all, and thus, different scoring rules should be used to get a better understanding of the predictive performance of forecasts (see, e.g., Scheuerer and Hamill 2015b; Taillardat 2021; Bjerregård et al. 2021). With that in mind, Scheuerer and Hamill (2015b) "strongly recommend that several different scores be always considered before drawing conclusions." This amplifies the need for numerous complementary proper scoring rules that are well-understood to facilitate forecast verification. In that direction, Dorninger et al. (2018) states that: "gaining an in-depth understanding of forecast performance depends on grasping the full meaning of the verification results." Interpretability of proper scoring rules can arise from being induced by a consistent scoring function for a functional (e.g., the squared error is induced by a scoring function consistent for the mean; Gneiting 2011), knowing what aspects of the forecast the scoring rule discriminates (e.g., the Dawid-Sebastiani score only discriminates forecasts through their mean and variance; Dawid and Sebastiani 1999) or knowing the limitations of a certain proper scoring rule (e.g., the variogram score is incapable of discriminating two forecasts that only differ by a constant bias; Scheuerer and Hamill 2015b).

In this context, interpretable proper scoring rules become verification methods of choice as the ranking of forecasts they produce can be more informative than the ranking of a more complex but less interpretable scoring rule. Section 4.2 provides an in-depth explanation of this in the case of univariate scoring rules. It is worth noting that interpretability of a scoring rule can also arise from its decomposition into meaningful terms (see, e.g., Bröcker 2009). This type of interpretability can be used complementarily to the framework proposed in this article.

Scheuerer and Hamill (2015b) proposed the variogram score to target the verification of the dependence structure. The variogram score of order p ($p > 0$) is defined as

$$\mathrm{VS}_p(F, \boldsymbol{y}) = \sum_{i,j=1}^{d} w_{ij} \left(\mathbb{E}_F \left[|X_i - X_j|^p \right] - |y_i - y_j|^p \right)^2,$$

where X_i is the i-th component of the random vector $\boldsymbol{X} \in \mathbb{R}^d$ following F, the w_{ij} are non-negative weights and $\boldsymbol{y} \in \mathbb{R}^d$ is an observation. The construction of the variogram score relies on two main principles. First, the variogram score is the weighted sum of scoring rules acting on the distribution of $\boldsymbol{X}_{i,j} = (X_i, X_j)$ and on paired components of the observations $y_{i,j}$. This *aggregation* principle allows the combination of proper scoring rules and summarizes them into a proper scoring rule acting on the whole distribution F and observations $\boldsymbol{y}$. Second, the scoring rules composing the weighted sum can be seen as a standard proper scoring rule applied to transformations of both forecasts and observations. Let us denote $\gamma_{i,j} : \boldsymbol{x} \mapsto |x_i - x_j|^p$ the transformation related to the variogram of order p, then the variogram score can be rewritten as

$$\mathrm{VS}_p(F, \boldsymbol{y}) = \sum_{i,j=1}^{d} w_{ij} \mathrm{SE}(\gamma_{i,j}(F), \gamma_{i,j}(\boldsymbol{y})),$$

where $\mathrm{SE}(F, y) = (\mathbb{E}_F[X] - y)^2$ is the univariate squared error and $\gamma_{i,j}(F)$ is the distribution of $\gamma_{i,j}(\boldsymbol{X})$ for $\boldsymbol{X}$ following F. This second principle is the *transformation* principle, allowing to build transformation-based proper scoring rules that can benefit from interpretability arising from a transformation (here, the variogram transformation $\gamma_{i,j}$) and the simplicity and interoperability of the proper scoring rule they rely on (here, the squared error).

We review the univariate and multivariate proper scoring rules through the lens of interpretability and by mentioning their known benefits and limitations. We formalize these two principles of aggregation and transformation to construct interpretable proper scoring rules for multivariate forecasts. To illustrate the use of these principles, we provide examples of transformation-and-aggregation-based scoring rules from both the literature on probabilistic forecast verification and quantities of interest. We conduct a simulation study to empirically demonstrate how transformation-and-aggregation-based scoring rules can be used. Additionally, we show how the aggregation and transformation principle can help bridging the gap between the proper scoring rules framework and the spatial verification tools (Gilleland et al., 2009; Dorninger et al., 2018).

The remainder of this article is organized as follows. Section 4.2 gives a general review of verification methods for univariate and multivariate forecasts. Section 4.3 introduces the framework of proper scoring rules based on transformation and aggregation for multivariate forecasts. Section 4.4 provides examples of transformation-and-aggregation-based scoring rules, including examples from the literature. Then, Section 4.5 showcases through different simulation setups how the framework proposed in this article can help build interpretable proper scoring rules. Finally, Section 4.6 provides a summary as well as a discussion on the verification of multivariate forecasts. Throughout the article, we focus on spatial forecasts for simplicity. However, the points made remain valid for any multivariate forecasts, including temporal forecasts or spatio-temporal forecasts.

4.2 Overview of verification tools for univariate and multivariate forecasts

This section presents the zoology of available verification tools and briefly summarizes their benefits and limitations. First, we define scoring rules and their key properties. Then, we recall univariate scoring rules, starting with ones derived from scoring functions used in point forecasting. Finally, we provide an overview of verification tools for multivariate forecasts.

4.2.1 Calibration, sharpness, and propriety

Gneiting et al. (2007) proposed a paradigm for the evaluation of probabilistic forecasts: "maximizing the sharpness of the predictive distributions subject to calibration". *Calibration* is the statistical compatibility between the forecast and the observations. *Sharpness* is the concentration of the forecast and is a property of the forecast itself. In other words, the paradigm aims at minimizing the uncertainty of the forecast given that the forecast is statistically consistent with the observations. Tsyplakov (2011) states that the notion of calibration in the paradigm is too vague but it holds if the definition of calibration is refined. This principle for the evaluation of probabilistic forecasts has reached a consensus in the field of probabilistic forecasting (see, e.g., Gneiting and Katzfuss 2014; Thorarinsdottir and Schuhen 2018). The paradigm proposed in Gneiting et al. (2007) is not the first mention of the link between sharpness and calibration: for example, Murphy and Winkler (1987) mentioned the relation between refinement (i.e., sharpness) and calibration.

For univariate forecasts, multiple definitions of calibration are available depending on the setting. The most used definition is *probabilistic calibration* and, broadly speaking, consists of computing the rank of observations among samples of the forecast and checking for uniformity with respect to observations. If the forecast is calibrated, observations should not be distinguishable from forecast samples, and thus, the distribution of their ranks should be uniform. Probabilistic calibration can be assessed by probability integral transform (PIT) histograms (Dawid, 1984) or rank histograms (Anderson, 1996; Talagrand et al., 1997) for ensemble forecasts when observations are stationary (i.e., their distribution is the same across time). The shape of the PIT or rank histogram gives information about the type of (potential) miscalibration: a triangular-shaped histogram suggests that the probabilistic forecast has a systematic bias, a $\cup$-shaped histogram suggests that the probabilistic forecast is under-dispersed and a $\cap$-shaped histogram suggests that the probabilistic forecast is over-dispersed. Moreover, probabilistic calibration implies that rank histograms should be uniform but uniformity is not sufficient. For example, rank histograms should also be uniform conditionally on different forecast scenarios (e.g., conditionally on the value of the observations available when the forecast is issued). Additionally, under certain hypotheses, calibration tools have been developed to consider real-world limitations such as serial dependence (Bröcker and Ben Bouallègue, 2020). Statistical tests have been developed to check the uniformity of rank histograms (Jolliffe and Primo, 2008). Readers interested in a more in-depth understanding of univariate forecast calibration are encouraged to consult Tsyplakov (2013, 2020).

For multivariate forecasts, a popular approach relies on a similar principle: first, multivariate forecast samples are transformed into univariate quantities using so-called pre-rank functions and then the calibration is assessed by techniques used in the univariate case (see, e.g., Gneiting et al. 2008). Pre-rank functions may be interpretable and allow targeting the calibration of specific aspects of the forecast such as the dependence structure. Readers interested in the calibration of multivariate forecasts can refer to Allen et al. (2024) for a comprehensive review of multivariate calibration.

A scoring rule S assigns a real-valued quantity $\mathrm{S}(F, y)$ to a forecast-observation pair (F, y), where $F \in \mathcal{F}$ is a probabilistic forecast and $\boldsymbol{y} \in \mathbb{R}^d$ is an observation. In the negative-oriented convention, a scoring rule S is *proper relative to the class* $\mathcal{F}$ if

$$\mathbb{E}_G[\mathrm{S}(G, \boldsymbol{Y})] \leq \mathbb{E}_G[\mathrm{S}(F, \boldsymbol{Y})] \tag{4.1}$$

for all $F, G \in \mathcal{F}$, where $\mathbb{E}_G[\cdots]$ is the expectation with respect to $\boldsymbol{Y} \sim G$. In simple terms, a scoring rule is proper relative to a class of distribution if its expected value is minimal when the true distribution is predicted, for any distribution within the class. Forecasts minimizing the expected scoring rule are said to be *efficient* and the other forecasts are said to be *sub-efficient*. Moreover, the scoring rule S is *strictly proper relative to the class* $\mathcal{F}$ if the equality in (4.1) holds if and only if $F = G$. This ensures the characterization of the ideal forecast (i.e., there is a unique efficient forecast and it is the true distribution). Moreover, proper scoring rules are powerful tools as they allow the assessment of calibration and sharpness simultaneously (Winkler, 1977; Winkler et al., 1996). Sharpness can be assessed individually using the entropy associated with proper scoring rules, defined by $e_{\mathrm{S}}(F) = \mathbb{E}_F[\mathrm{S}(F, \boldsymbol{Y})]$. The sharper the forecast, the smaller its entropy. Strictly proper scoring rules can also be used to infer the parameters of a parametric probabilistic forecast (see, e.g., Gneiting et al. 2005; Pacchiardi et al. 2024).

Regardless of all the interesting properties of proper scoring rules, it is worth noting that they have some limitations. Proper scoring rules may have multiple efficient forecasts (i.e., associated with their minimal expected value) and, in the general setting, no guarantee is given on their relevance. Moreover, strict propriety ensures that the efficient forecast is unique and that it is the ideal forecast (i.e., the true distribution), however, no guarantee is available for forecasts within the vicinity of the minimum in the general case. This is particularly problematic since, in practice, the unavailability of the ideal distribution makes it impossible to know if the minimum expected score is achieved. In the case of calibrated forecasts, the expected scoring rule is the entropy of the forecast and the ranking of forecasts is thus linked to the information carried by the forecast (see Corollary 4, Holzmann and Eulert 2014 for the complete result). These limitations may explain the plurality of scoring rules depending on application fields.

4.2.2 Univariate scoring rules

We recall classical univariate scoring rules to explain key concepts. Some univariate scoring rules will be useful for the multivariate scoring rules construction framework proposed in Section 4.3. Let $\mathcal{P}(E)$ denote the class of Borel probability measures on E. We consider $F \in \mathcal{F} \subseteq \mathcal{P}(\mathbb{R})$ a probabilistic forecast in the form of its cumulative distribution function (cdf) and $y \in \mathbb{R}$ an observation. When the probabilistic forecast F has a probability density function (pdf), it will be denoted f.

The simplest scoring rules can be derived from scoring functions used to assess point forecasts. The squared error (SE) is the most popular and is known through its averaged value (the mean squared error; MSE) or the square root of its average (the root mean squared error; RMSE) which has the advantage of being expressed in the same units as the observations. As a scoring rule, the SE is expressed as

$$\mathrm{SE}(F, y) = (\mu_F - y)^2, \tag{4.2}$$

where μ_F denotes the mean of the predicted distribution F. The SE solely discriminates the mean of the forecast (see Appendix 4.7.1); efficient forecasts for SE are the ones matching the mean of the true distribution. The SE is proper relative to $\mathcal{P}_2(\mathbb{R})$, the class of Borel probability measures on $\mathbb{R}$ with a finite second moment (i.e., finite variance). Note that the SE cannot be strictly proper as the equality of mean does not imply the equality of distributions.

Another well-known scoring rule is the absolute error (AE) defined by

$$\mathrm{AE}(F, y) = |\mathrm{med}(F) - y|, \tag{4.3}$$

where $\mathrm{med}(F)$ is the median of the predicted distribution F. The mean absolute error (MAE), the average of the absolute error, is the most seen form of the AE and it is also expressed in the same units as the observations. Efficient forecasts are forecasts that have a median equal to the median of the true distribution. The AE is proper relative to $\mathcal{P}_1(\mathbb{R})$ but not strictly proper. Similarly, the quantile score (QS), also known as the pinball loss, is a scoring rule focusing on quantiles of level α defined by

$$\mathrm{QS}_\alpha(F, y) = (\mathbb{1}_{y \leq F^{-1}(\alpha)} - \alpha)(F^{-1}(\alpha) - y) \tag{4.4}$$

where $0 < \alpha < 1$ is a probability level and $F^{-1}(\alpha)$ is the predicted quantile of level α. The case $\alpha = 0.5$ corresponds to the AE up to a factor 2. The QS of level α is proper relative to $\mathcal{P}_1(\mathbb{R})$ but not strictly proper since efficient forecasts are ones correctly predicting the quantile of level α (see, e.g., Friederichs and Hense 2008).

Another summary statistic of interest is the exceedance of a threshold $t \in \mathbb{R}$. The Brier score (BS; Brier 1950) was initially introduced for binary predictions but allows also to discriminate forecasts based on the exceedance of a threshold t. For probabilistic forecasts, the BS is defined as

$$\mathrm{BS}_t(F, y) = ((1 - F(t)) - \mathbb{1}_{y>t})^2 = (F(t) - \mathbb{1}_{y \leq t})^2, \tag{4.5}$$

where $1 - F(t)$ is the predicted probability that the threshold t is exceeded. The BS is proper relative to $\mathcal{P}(\mathbb{R})$ but not strictly proper. Binary events (e.g., exceedance of thresholds) are relevant in weather forecasting as they are used, for example, in operational settings for decision-making.

All the scoring rules presented above are proper but not strictly proper since they only discriminate against specific summary statistics instead of the whole distribution. Nonetheless, they are still used as they allow forecasters to verify specific characteristics of the forecast: the mean, the median, the quantile of level α or the exceedance of a threshold t. The simplicity of these scoring rules makes them interpretable, thus making them essential verification tools.

Some univariate scoring rules contain a summary statistic: for example, the formulas of the QS (4.4) or the BS (4.5) contain the exceedance of a threshold t and the quantile of level α, respectively. They can be seen as a scoring function applied to a summary statistic. This duality can be understood through the link between scoring functions and scoring rules through consistent functionals as presented in Gneiting (2011) or Section 2.2 in Lerch et al. (2017).

Other summary statistics can be of interest depending on applications. Nonetheless, it is worth noting that mispecifications of numerous summary statistics cannot be discriminated because of their *non-elicitability*. Non-elicitability of a transformation implies that no proper scoring rule can be constructed such that efficient forecasts are forecasts where the transformation is equal to the one of the true distribution. For example, the variance is known to be non-elicitable; however, it is jointly elicitable with the mean (see, e.g., Brehmer 2017). Readers interested in details regarding elicitable, non-elicitable and jointly elicitable transformations may refer to Gneiting (2011), Brehmer and Strokorb (2019) and references therein.

A strictly proper scoring rule should discriminate the whole distribution and not only specific summary statistics. The continuous ranked probability score (CRPS; Matheson and Winkler 1976) is the most popular univariate scoring rule in weather forecasting applications and can

be expressed by the following expressions

$$\text{CRPS}(F, y) = \mathbb{E}_F|X - y| - \frac{1}{2}\mathbb{E}_F|X - X'|, \tag{4.6}$$

$$= \int_{\mathbb{R}} \text{BS}_z(F, y)\mathrm{d}z, \tag{4.7}$$

$$= 2\int_0^1 \text{QS}_\alpha(F, y)\mathrm{d}\alpha, \tag{4.8}$$

where $y \in \mathbb{R}$ and X and X' are independent random variables following F, with a finite first moment. Equations (4.7) and (4.8) show that the CRPS is linked with the BS and the QS. Broadly speaking, as the QS discriminates a quantile associated with a specific level, integrating the QS across all levels discriminates the quantile function that fully characterizes univariate distributions. Similarly, integrating the BS across all thresholds discriminates the cumulative distribution function that also fully characterizes univariate distributions. The CRPS is a strictly proper scoring rule relative to $\mathcal{P}_1(\mathbb{R})$, the class of Borel probability measures on $\mathbb{R}$ with a finite first moment. In addition, Equation (4.6) indicates the CRPS values have the same units as observations. In the case of deterministic forecasts, the CRPS reduces to the absolute error, in its scoring function form (Hersbach, 2000). The use of the CRPS for ensemble forecast is straightforward using expectations as in (4.6). Ferro et al. (2008) and Zamo and Naveau (2017) studied estimators of the CRPS for ensemble forecasts.

In addition to scoring rules based on scoring functions, some scoring rules use the moments of the probabilistic forecast F. The SE (4.2) depends on the forecast only through its mean μ_F. The Dawid-Sebastiani score (DSS; Dawid and Sebastiani 1999) is a scoring rule depending on the forecast F only through its first two central moments. The DSS is expressed as

$$\text{DSS}(F, y) = 2\log(\sigma_F) + \frac{(\mu_F - y)^2}{{\sigma_F}^2}, \tag{4.9}$$

where μ_F and ${\sigma_F}^2$ are the mean and the variance of the distribution F. The DSS is proper relative to $\mathcal{P}_2(\mathbb{R})$ but not strictly proper, since efficient forecasts only need to correctly predict the first two central moments (see Appendix 4.7.1). Dawid and Sebastiani (1999) proposed a more general class of proper scoring rules but the DSS, as defined in (4.9), can be seen as a special case of the logarithmic score (up to an additive constant), introduced further down.

Another scoring rule relying on the central moments of the probabilistic forecast F up to order three is the error-spread score (ESS; Christensen et al. 2014). The ESS is defined as

$$\text{ESS}(F, y) = ({\sigma_F}^2 - (\mu_F - y)^2 - (\mu_F - y)\sigma_F\gamma_F)^2, \tag{4.10}$$

where μ_F, σ_F^2 and γ_F are the mean, the variance and the skewness of the probabilistic forecast F. The ESS is proper relative to $\mathcal{P}_4(\mathbb{R})$. As for the other scoring rules only based on moments of the forecast presented above, the expected ESS compares the probabilistic forecast F with the true distribution only via their four first moments (see Appendix 4.7.1). Scoring rules based on central moments of higher order could be built following the process described in Christensen et al. (2014). Such scoring rules would benefit from the interpretability induced by their construction and the ease to be applied to ensemble forecasts. However, they would also inherit the limitation of being only proper.

When the probabilistic forecast F has a pdf f, scoring rules of a different type can be defined. Let $\mathcal{L}_\alpha(\mathbb{R})$ denote the class of probability measures on $\mathbb{R}$ that are absolutely continuous with respect to μ (usually taken as the Lebesgue measure) and have μ-density f such that

$$\|f\|_\alpha = \left(\int_{\mathbb{R}} f(x)^\alpha \mu(\mathrm{d}x)\right)^{1/\alpha} < \infty.$$

The most popular scoring rule based on the pdf is the logarithmic score (also known as ignorance score; Good 1952; Roulston and Smith 2002). The logarithmic score is defined as

$$\text{LogS}(F, y) = -\log(f(y)), \tag{4.11}$$

for y such that $f(y) > 0$. In its formulation, the logarithmic score is different from the scoring rules seen previously. Good (1952) proposed the logarithmic score knowing its link with the theory of information: its entropy is the Shannon entropy (Shannon, 1948) and its expectation is related to the Kullback-Leibler divergence (Kullback and Leibler, 1951) (see Appendix 4.7.1). The logarithmic score is strictly proper relative to the class $\mathcal{L}_1(\mathbb{R})$. Moreover, inference via minimization of the expected logarithmic score is equivalent to maximum likelihood estimation (see, e.g., Dawid et al. 2015). The logarithmic score belongs to the family of *local scoring rules*, which are scoring rules only depending on y, $f(y)$ and its derivatives up to a finite order. Another local scoring rule is the Hyvärinen score (also known as the gradient scoring rule; Hyvärinen 2005) and it is defined as

$$\text{HS}(F, y) = 2\frac{f''(y)}{f(y)} - \frac{f'(y)^2}{f(y)^2},$$

for y such that $f(y) > 0$. The Hyvärinen score is proper relative to the subclass of $\mathcal{P}(\mathbb{R})$ such that the density f exists, is twice continuously differentiable and satisfies $f'(x)/f(x) \to 0$ as $|x| \to \infty$. It is worth noticing that the Hyvärinen score can be computed even if f is only known up to a scale factor (e.g., up to a normalizing constant). This property allows circumventing the use of Monte Carlo methods or approximations of the normalizing constant when it is unavailable or hard to compute. This is a property of local proper scoring rules except for the logarithmic score (Parry et al., 2012). Readers eager to learn more about local proper scoring rules may refer to Parry et al. (2012) and Ehm and Gneiting (2012).

The logarithmic score and the Hyvärinen score do not allow f to be zero. To overcome this limitation, scoring rules expressed in terms of the L_α-norm have been proposed. The quadratic score is defined as

$$\text{QuadS}(F, y) = \|f\|_2^2 - 2f(y),$$

where $\|f\|_2^2 = \int_{\mathbb{R}} f(y)^2 \mathrm{d}y$. The quadratic score is strictly proper relative to the class $\mathcal{L}_2(\mathbb{R})$.

The pseudospherical score is defined as

$$\text{PseudoS}(F, y) = -f(y)^{\alpha-1}/\|f\|_\alpha^{\alpha-1},$$

with $\alpha > 1$. For $\alpha = 2$, it reduces to the spherical score (see, e.g., Jose 2007). The pseudospherical score is strictly proper relative to the class $\mathcal{L}_\alpha(\mathbb{R})$. The four scoring rules presented above have been criticized as they do not encourage a high probability in the vicinity of the observation y (Gneiting and Raftery, 2007). In particular, as the logarithmic score is more sensitive to outliers, probabilistic forecasts inferred by its minimization may be overdispersive (Gneiting et al., 2005). Moreover, the pdf is not always available, for example in the case of ensemble forecasts.

Readers may refer to the various reviews of scoring rules available (see, e.g., Bröcker and Smith 2007; Gneiting and Raftery 2007; Gneiting and Katzfuss 2014; Thorarinsdottir and Schuhen 2018; Alexander et al. 2022). Formulas of the expected scoring rules presented are available in Appendix 4.7.1.

Strictly proper scoring rules can be seen as more powerful than proper scoring rules. This is theoretically true when the interest is in identifying the ideal forecast (i.e., the true distribution). Regardless, in practice, scoring rules are also used to rank probabilistic forecasts and with that in mind, a given ranking of forecasts in terms of the expectation of a strictly proper scoring rule (such as the CRPS) is harder to interpret than a ranking in terms of the

expectation of a proper but more interpretable scoring rule (such as the SE). The SE is known to discriminate the mean, and thus, a better rank in terms of expected SE implies a better prediction of the mean of the true distribution. Conversely, a better ranking in terms of CRPS implies a better prediction of the whole prediction, but it might not be useful as is, and other verification tools will be needed to know what caused this ranking. When forecasts are not calibrated, there seems to be a trade-off between interpretability and discriminatory power and this becomes more prominent in a multivariate setting. However, simpler interpretable tools and discriminatory-powerful tools can be used complementarily. The framework proposed in Section 4.3 aims at helping the construction of interpretable proper scoring rules.

4.2.3 Multivariate scoring rules

In a multivariate setting, forecasters cannot solely use univariate scoring rules as they are not able to discriminate forecasts beyond their 1-dimensional marginals. Univariate scoring rules cannot discriminate the dependence structure between the univariate margins. Multivariate forecasts can be applied in different setups: spatial forecasts, temporal forecasts, multivariable forecasts or any combination of these categories (e.g., spatio-temporal forecasts of multiple variables). Considering weather forecasting, a spatial forecast could aim at predicting temperatures across multiple locations. A temporal forecast could be focused on predicting rainfall at multiple lead times at a given location. A multivariable forecast could predict both eastward and northward components of the wind. In the following, we consider $F \in \mathcal{F} \subseteq \mathcal{P}(\mathbb{R}^d)$ a multivariate probabilistic forecast and $\boldsymbol{y} \in \mathbb{R}^d$ an observation.

Even if there is no natural ordering in the multivariate case, the notions of median and quantile can be adapted using level sets, and then scoring rules using these quantities can be constructed (see, e.g., Meng et al. 2023). Nonetheless, as the mean is well-defined, the squared error (SE) can be defined in the multivariate setting :

$$\mathrm{SE}(F, \boldsymbol{y}) = \|\boldsymbol{\mu}_F - \boldsymbol{y}\|_2^2, \tag{4.12}$$

where μ_F is the mean vector of the distribution F. Similar to the univariate case, the SE is proper relative to $\mathcal{P}_2(\mathbb{R}^d)$. Moments are well-defined in the multivariate case allowing the multivariate version of the Dawid-Sebastiani score to be defined. The Dawid-Sebastiani score (DSS) was proposed in Dawid and Sebastiani (1999) as

$$\mathrm{DSS}(F, \boldsymbol{y}) = \log(\det \Sigma_F) + (\boldsymbol{\mu}_F - \boldsymbol{y})^T \Sigma_F^{-1} (\boldsymbol{\mu}_F - \boldsymbol{y}),$$

where $\boldsymbol{\mu}_F$ and Σ_F are the mean vector and the covariance matrix of the distribution F. The DSS is proper relative to $\mathcal{P}_2(\mathbb{R}^d)$ and it becomes strictly proper relative to any convex class of probability measures characterized by their first two moments (Gneiting and Raftery, 2007). The second term in the DSS is the squared Mahalanobis distance between $\boldsymbol{y}$ and $\boldsymbol{\mu}_F$.

To define a strictly proper scoring rule for multivariate forecast, Gneiting and Raftery (2007) proposed the energy score (ES) as a generalization of the CRPS to the multivariate case. The ES is defined by

$$\mathrm{ES}_\alpha(F, \boldsymbol{y}) = \mathbb{E}_F \|\boldsymbol{X} - \boldsymbol{y}\|_2^\alpha - \frac{1}{2} \mathbb{E}_F \|\boldsymbol{X} - \boldsymbol{X}'\|_2^\alpha, \tag{4.13}$$

where $\alpha \in (0, 2)$ and $F \in \mathcal{P}_\alpha(\mathbb{R}^d)$, the class of Borel probability measures on $\mathbb{R}^d$ such that the moment of order α is finite. The definition of the ES is related to the kernel form of the CRPS (4.6), to which the ES reduces for $d = 1$ and $\alpha = 1$. As pointed out in Gneiting and Raftery (2007), in the limiting case $\alpha = 2$, the ES becomes the SE (4.12). The ES is strictly proper relative to $\mathcal{P}_\alpha(\mathbb{R}^d)$ (Székely, 2003; Gneiting and Raftery, 2007) and is suited for ensemble forecasts (Gneiting et al., 2008). Moreover, the parameter α gives some flexibility: a small value

of α can be chosen and still lead to a strictly proper scoring rule, for example, when higher-order moments are ill-defined. The discrimination ability of the ES has been studied in numerous studies (see, e.g., Pinson and Girard 2012; Pinson and Tastu 2013; Scheuerer and Hamill 2015b). Pinson and Girard (2012) studied the ability of the ES to discriminate among rival sets of scenarios (i.e., forecasts) of wind power generation. In the case of bivariate Gaussian processes, Pinson and Tastu (2013) illustrated that the ES appears to be more sensitive to misspecifications of the mean rather than misspecifications of the variance or dependence structure. The lack of sensitivity to misspecifications of the dependence structure has been confirmed in Scheuerer and Hamill (2015b) using multivariate Gaussian random vectors of higher dimension. Moreover, the discriminatory power of the ES deteriorates in higher dimensions (Pinson and Tastu, 2013).

To overcome the discriminatory limitation of the ES, Scheuerer and Hamill (2015b) proposed the variogram score (VS), a score targeting the verification of the dependence structure. The VS of order p is defined as

$$\mathrm{VS}_p(F, \boldsymbol{y}) = \sum_{i,j=1}^{d} w_{ij} \left(\mathbb{E}_F \left[|X_i - X_j|^p \right] - |y_i - y_j|^p \right)^2 \tag{4.14}$$

where X_i is the i-th component of the random vector X following F, w_{ij} are nonnegative weights and $p > 0$. The variogram score capitalizes on the variogram, used in spatial statistics to access the dependence structure. The VS cannot detect an equal bias across all components. The VS of order p is proper relative to the class of Borel probability measures on $\mathbb{R}^d$ such that the $2p$-th moments of all univariate margins are finite. The weights w_{ij} can be selected to emphasize or depreciate certain pair interactions. For example, in a spatial context, it can be expected the dependence between pairs decays with the distance: choosing the weights proportional to the inverse of the distance between locations can increase the signal-to-noise ratio and improve the discriminatory power of the VS (Scheuerer and Hamill, 2015b).

When the pdf f of the probabilistic forecast F is available, multivariate versions of the univariate scoring rules based on the pdf are available. The multivariate versions of the scoring rules have the same properties and limitations as their univariate counterpart. The logarithmic score (4.11) has a natural multivariate version :

$$\mathrm{LogS}(F, \boldsymbol{y}) = -\log(f(\boldsymbol{y})),$$

for $\boldsymbol{y}$ such that $f(\boldsymbol{y}) > 0$. The logarithmic score is strictly proper relative to the class $\mathcal{L}_1(\mathbb{R}^d)$.

The Hyvärinen score (HS; Hyvärinen 2005) was initially proposed in its multivariate form

$$\mathrm{HS}(F, \boldsymbol{y}) = 2\Delta \log(f(\boldsymbol{y})) + |\nabla \log(f(\boldsymbol{y}))|^2,$$

for $\boldsymbol{y}$ such that $f(\boldsymbol{y}) > 0$, where Δ is the Laplace operator (i.e., the sum of the second-order partial derivatives) and ∇ is the gradient operator (i.e., vector of the first-order partial derivatives). In the multivariate setting, the HS can also be computed if the predicted pdf is known up to a normalizing constant. The HS is proper relative to the subclass of $\mathcal{P}(\mathbb{R}^d)$ such that the density f exists, is twice continuously differentiable and satisfies $\|\nabla \log(f(x))\| \to 0$ as $\|x\| \to \infty$.

The quadratic score and pseudospherical score are directly suited to the multivariate setting :

$$\begin{aligned} \mathrm{QuadS}(F, \boldsymbol{y}) &= \|f\|_2^2 - 2f(\boldsymbol{y}); \\ \mathrm{PseudoS}(F, \boldsymbol{y}) &= -f(\boldsymbol{y})^{\alpha-1} / \|f\|_\alpha^{\alpha-1}, \end{aligned}$$

where $\|f\|_\alpha = \left(\int_{\mathbb{R}^d} f(\boldsymbol{y})^\alpha \mathrm{d}\boldsymbol{y}\right)^{1/\alpha}$. The quadratic score is strictly proper relative to the class $\mathcal{L}_2(\mathbb{R}^d)$. The pseudospherical score is strictly proper relative to the class $\mathcal{L}_\alpha(\mathbb{R}^d)$.

Additionally, other multivariate scoring rules have been proposed among which the marginal-copula score (Ziel and Berk, 2019) or wavelet-based scoring rules (see, e.g., Buschow et al. 2019). These scoring rules will be briefly mentioned in Section 4.4 in light of the proper scoring rule construction framework proposed in this article. Appendix 4.7.2 provides formulas for the expected multivariate scoring rules presented above.

4.2.4 Spatial verification tools

Spatial forecasts are a very important group of multivariate forecasts as they are involved in various applications (e.g., weather or renewable energy forecasting). Spatial fields are often characterized by high dimensionality and potentially strong correlations between neighboring locations. These characteristics make the verification of spatial forecasts very demanding in terms of discriminating misspecified dependence structures, for example. In the case of spatial forecasts, it is known that traditional verification methods (e.g., gridpoint-by-gridpoint verification) may result in a double penalty. The *double-penalty effect* was pinned in Ebert (2008) and refers to the fact that if a forecast presents a spatial (or temporal) shift with respect to observations, the error made would be penalized twice: once where the event was observed and again where the forecast predicted it. In particular, high-resolution forecasts are more penalized than less realistic blurry forecasts. The double-penalty effect may also affect spatio-temporal forecasts in general.

In parallel with the development of scoring rules, various application-focused spatial verification methods have been developed to evaluate weather forecasts. The efforts toward improving spatial verification methods have been guided by two projects: the intercomparison project (ICP; Gilleland et al. 2009) and its second phase, called Mesoscale Verification Intercomparison over Complex Terrain (MesoVICT; Dorninger et al. 2018). These projects resulted in the comparison of spatial verification methods with a particular focus on understanding their limitations and clarifying their interpretability. Only a few links exist between the approaches studied in these projects (and the work they induced) and the proper scoring rules framework. In particular, Casati et al. (2022) noted "a lack of representation of novel spatial verification methods for ensemble prediction systems". In general, there is a clear lack of methods focusing on the spatial verification of probabilistic forecasts. Moreover, to help bridging the gap between the two communities, we would like to recall the approach of spatial verification tools in the light of the scoring rule framework introduced above.

One of the goals of the ICP was to provide insights on how to develop methods robust to the double-penalty effect. In particular, Gilleland et al. (2009) proposed a classification of spatial verification tools updated later in Dorninger et al. (2018) resulting in a five-category classification. The classes differ in the computing principle they rely on. Not all spatial verification tools mentioned in these studies can be applied to probabilistic forecasts, some of them can solely be applied to deterministic forecasts. In the following description of the classes, we try to focus on methods suited to probabilistic forecasts or at least the special case of ensemble forecasts.

Neighborhood-based methods consist of applying a smoothing filter to the forecast and observation fields to prevent the double-penalty effect. The smoothing filter can take various forms (e.g., a minimum, a maximum, a mean, or a Gaussian filter) and be applied over a given neighborhood. For example, Stein and Stoop (2022) proposed a neighborhood-based CRPS for ensemble forecasts gathering forecasts and observations made within the neighborhood of the location considered. The use of a neighborhood prevents the double-penalty effect from taking place at scales smaller than that of the neighborhood. In this general definition, neighborhood-based methods can lead to proper scoring rules, in particular, see the notion of patches in Section 4.4.

Scale-separation techniques denote methods for which the verification is obtained after comparing forecast and observation fields across different scales. The scale-separation process can be seen as several single-bandpass spatial filters (e.g., projection onto a base of wavelets as wavelet-based scoring rules; Buschow et al. 2019). However, in order to obtain proper scoring rules, the comparison of the scale-specific characteristics needs to be performed using a proper scoring rule. Section 4.4 provides a discussion on wavelet-based scoring rules and their propriety.

Object-based methods rely on the identification of objects of interest and the comparison of the objects obtained in the forecast and observation fields. Object identification is application-dependent and can take the form of objects that forecasters are familiar with (e.g., storm cells for precipitation forecasts). A well-known verification tool within this class is the structure-amplitude-location (SAL; Wernli et al. 2008) method which has been generalized to ensemble forecasts in Radanovics et al. (2018). The three components of the ensemble SAL do not lead to proper scoring rules. They rely on the mean of the forecast within scoring functions inconsistent with the mean. Thus, the ideal forecast does not minimize the expected value. Nonetheless, the three components of the SAL method could be adapted to use proper scoring rules sensitive to the misspecification of the same features.

Field-deformation techniques consist of deforming the forecasts field into the observation field (the similarity between the fields can be ensured by a metric of interest). The field of distortion associated with the morphing of the forecast field into the observation field becomes a measure of the predictive performance of the forecast (see, e.g., Han and Szunyogh 2018).

Distance measures between binary images, such as exceedance of a threshold of interest, of the forecast and observation fields. These methods are inspired by development in image processing (e.g., Baddeley's delta measure Gilleland 2011).

These five categories are partially overlapping as it can be argued that some methods belong to multiple categories (e.g., some distance measures techniques can be seen as a mix of field-deformation and object-based). They define different principles that can be used to build verification tools that are not subject to the double-penalty effect. The reader may refer to Dorninger et al. (2018) and references therein for details on the classification and the spatial verification methods not used thereafter. The frontier between the aforementioned spatial verification methods and the proper scoring rules framework is porous with, for example, wavelet-based scoring rules belonging to both. It appears that numerous spatial verification methods seek interpretability and we believe that this is not incompatible with the use of proper scoring rules. We propose the following framework to facilitate the construction of interpretable proper scoring rules.

4.3 A framework for interpretable proper scoring rules

We define a framework to design proper scoring rules for multivariate forecasts. Its definition is motivated by remarks on the multivariate forecasts literature and operational use. There seems to be a growing consensus around the fact that no single verification method has it all (see, e.g., Bjerregård et al. 2021). Most of the studies comparing forecast verification methods highlight that verification procedures should not be reduced to the use of a single method and that each procedure needs to be well suited to the context (see, e.g., Scheuerer and Hamill 2015b; Thorarinsdottir and Schuhen 2018). Moreover, from a more theoretical point of view, (strict) propriety does not ensure discrimination ability and different (strictly) proper scoring rules can lead to different rankings of sub-efficient forecasts.

Standard verification procedures gradually increase the complexity of the quantities verified. Procedures often start by verifying simple quantities such as quantiles, mean, or binary events (e.g., prediction of dry/wet events for precipitation). If multiple forecasts have a satisfying performance for these quantities, marginal distributions of the multivariate forecast can

be verified using univariate scoring rules. Finally, multivariate-related quantities, such as the dependence structure, can be verified through multivariate scoring rules. Forecasters rely on multiple verification methods to evaluate a forecast and ideally, the verification method should be interpretable by targeting specific aspects of the distribution or thanks to the forecaster's experience. This type of verification procedure allows the forecaster to understand what characterizes the predictive performance of a forecast instead of directly looking at a strictly proper scoring rule giving an encapsulated summary of the predictive performance.

Various multivariate forecast calibration methods rely on the calibration of univariate quantities obtained by dimension reduction techniques. As the general principle of multivariate calibration leans on studying the calibration of quantities obtained by pre-rank functions, Allen et al. (2024) argue that calibration procedures should not rely on a single pre-rank function and should instead use multiple simple pre-rank functions and leverage the interpretability of the PIT/rank histograms associated. A similar principle can be applied to increase the interpretability of verification methods based on scoring rules.

As general multivariate strictly proper scoring rules fail to discriminate forecasts with respect to arbitrary misspecifications and they may lead to different ranking of sub-efficient forecasts, multivariate verification could benefit from using multiple proper scoring rules targeting specific aspects of the forecasts. Thereby, forecasters know which aspect of the observations are well-predicted by the forecast and can update their forecast or select the best forecast among others in the light of this better understanding of the forecast. To facilitate the construction of interpretable proper scoring rules, we define a framework based on two principles: transformation and aggregation.

The transformation principle consists of transforming both forecast and observation before applying a scoring rule. Heinrich-Mertsching et al. (2024) introduced this general principle in the context of point processes. In particular, they present scoring rules based on summary statistics targeting the clustering behavior or the intensity of the processes. In a more general context, the use of transformations was disseminated in the literature for several years (see Section 4.4). Proposition 4.1 shows how transformations can be used to construct proper scoring rules.

Proposition 4.1. *Let $\mathcal{F} \subset \mathcal{P}(\mathbb{R}^d)$ be a class of Borel probability measure on $\mathbb{R}^d$ and let $F \in \mathcal{F}$ be a forecast and $\boldsymbol{y} \in \mathbb{R}^d$ an observation. Let $T : \mathbb{R}^d \to \mathbb{R}^k$ be a transformation and let* S *be a scoring rule on $\mathbb{R}^k$ that is proper relative to $T(\mathcal{F}) = \{\mathcal{L}(T(\boldsymbol{X})), \boldsymbol{X} \sim F \in \mathcal{F}\}$. Then, the scoring rule*

$$\mathrm{S}_T(F, \boldsymbol{y}) = \mathrm{S}(T(F), T(\boldsymbol{y}))$$

is proper relative to $\mathcal{F}$. If S *is strictly proper relative to $T(\mathcal{F})$ and T is injective, then the resulting scoring rule S_T is strictly proper relative to $\mathcal{F}$.*

To gain interpretability, it is natural to have dimension-reducing transformations (i.e., $k < d$), which generally leads to T not being injective and S_T not being strictly proper. Nonetheless, as expressed previously, interpretability is important and it can mostly be leveraged if the transformation simplifies the multivariate quantities. Particularly, it is generally preferred to choose $k = 1$ to make the quantity easier to interpret and focus on specific information contained in the forecast or the observation. Straightforward transformations can be projections on a k-dimensional margin or a summary statistic relevant to the forecast type such as the total over a domain in the case of precipitations. Simple transformations may be preferred for their interpretability and their potential lack of discriminatory power can be made up for via the use of multiple simpler transformations. Numerous examples of transformations are presented, discussed, and linked to the literature in Section 4.4. The proof of Proposition 4.1 is provided in Appendix 4.7.3.

The second principle is the aggregation of scoring rules. Aggregation can be used on scoring rules in order to combine them and obtain a single scoring rule summarizing the evaluation. It can be used to operate on scoring rules acting on different spaces, times or locations. Note that Dawid and Musio (2014) introduced the notion of *composite score* which is related to the aggregation principle but is closer to the combined application of both principles. Proposition 4.2 presents a general aggregation principle to build proper scoring rules. This principle has been known since proper scoring rules have been introduced.

Proposition 4.2. *Let $\mathcal{S} = \{\mathrm{S}_i\}_{1\leq i\leq m}$ be a set of proper scoring rules relative to $\mathcal{F} \subset \mathcal{P}(\mathbb{R}^d)$. Let $\boldsymbol{w} = \{w_i\}_{1\leq i\leq m}$ be nonnegative weights. Then, the scoring rule*

$$\mathrm{S}_{\mathcal{S},\boldsymbol{w}}(F, \boldsymbol{y}) = \sum_{i=1}^{m} w_i \mathrm{S}_i(F, \boldsymbol{y})$$

is proper relative to $\mathcal{F}$. If at least one scoring rule S_i is strictly proper relative to $\mathcal{F}$ and $w_i > 0$, the aggregated scoring rule $\mathrm{S}_{\mathcal{S},\boldsymbol{w}}$ is strictly proper relative to $\mathcal{F}$.

It is worth noting that Proposition 4.2 does not specify any strict condition for the scoring rules used. For example, the scoring rules aggregated do not need to be the same or do not need to be expressed in the same units. Aggregated scoring rules can be used to summarize the evaluation of univariate probabilistic forecasts (e.g., aggregation of CRPS at different locations) or to summarize complementary scoring rules (e.g., aggregation of Brier score and a threshold-weighted CRPS). Unless stated otherwise, for simplicity, we will restrict ourselves to cases where the aggregated scoring rules are of the same type. Bolin and Wallin (2023) showed that the aggregation of scoring rules can lead to unintuitive behaviors. For the aggregation of univariate scoring rules, they showed that scoring rules do not necessarily have the same dependence on the scale of the forecasted phenomenon: this leads to scoring rules putting more (or less) emphasis on the forecasts with larger scales. They define and propose local scale-invariant scoring rules to make scale-agnostic scoring rules. When performing aggregation, it is important to be aware of potential preferences or biases of the scoring rules.

We only consider aggregation of proper scoring rules through a weighted sum. To conserve (strict) propriety of scoring rules, aggregations can take, more generally, the form of (strictly) isotonic transformations, such as a multiplicative structure when positive scoring rules are considered (Ziel and Berk, 2019).

The two principles of Proposition 4.1 and Proposition 4.2 can be used simultaneously to create proper scoring rules based on both transformations and aggregation as presented in Corollary 4.1.

Corollary 4.1. *Let $\mathcal{T} = \{T_i\}_{1\leq i\leq m}$ be a set of transformations from $\mathbb{R}^d$ to $\mathbb{R}^k$. Let $\mathcal{S}_{\mathcal{T}} = \{\mathrm{S}_{T_i}\}_{1\leq i\leq m}$ be a set of proper scoring rules where S is proper relative to $T_i(\mathcal{F})$, for all $1 \leq i \leq m$. Let $\boldsymbol{w} = \{w_i\}_{1\leq i\leq m}$ be nonnegative weights. Then, the scoring rule*

$$\mathrm{S}_{\mathcal{S}_{\mathcal{T}},\boldsymbol{w}}(F, \boldsymbol{y}) = \sum_{i=1}^{m} w_i \mathrm{S}_{T_i}(F, \boldsymbol{y})$$

is proper relative to $\mathcal{F}$.

Strict propriety relative to $\mathcal{F}$ of the resulting scoring rule is obtained as soon as there exists $1 \leq i \leq m$ such that S is strictly proper relative to $T_i(\mathcal{F})$, T_i is injective and $w_i > 0$. The result of Corollary 4.1 can be extended to transformations with images in different dimensions and paired with different scoring rules (see Appendix 4.7.4).

As we will see in the examples developed in the following section, numerous scoring rules used in the literature are based on these two principles of aggregation and transformation.

Decomposition of kernel scoring rules. We briefly discuss the link between the transformation and aggregation principles for scoring rules and the specific class of kernel scoring rules. A kernel on $\mathbb{R}^d$ is a measurable function $\rho : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ satisfying the following two properties:

i) (symmetry) $\rho(\boldsymbol{x}_1, \boldsymbol{x}_2) = \rho(\boldsymbol{x}_2, \boldsymbol{x}_1)$ for all $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{R}^d$;

ii) (non-negativity) $\sum_{1 \leq i \leq j \leq n} a_i a_j \rho(\boldsymbol{x}_i, \boldsymbol{x}_j) \geq 0$ for all $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n \in \mathbb{R}^d$ and $a_1, \dots, a_n \in \mathbb{R}$, for all $n \in \mathbb{N}$.

The kernel scoring rule S_ρ associated with the kernel ρ is defined on the space of predictive distributions

$$\mathcal{P}_\rho = \left\{ F \in \mathcal{P}(\mathbb{R}^d) \colon \int \sqrt{\rho(x,x)} F(\mathrm{d}x) < +\infty \right\}$$

by

$$\begin{aligned} \mathrm{S}_\rho(F, \boldsymbol{y}) &= \frac{1}{2} \int_{\mathbb{R}^d \times \mathbb{R}^d} \rho(x_1, x_2)(F - \delta_{\boldsymbol{y}})(\mathrm{d}x_1)(F - \delta_{\boldsymbol{y}})(\mathrm{d}x_2), \\ &= \frac{1}{2} \mathbb{E}_F[\rho(\boldsymbol{X}, \boldsymbol{X}')] + \frac{1}{2}\rho(\boldsymbol{y}, \boldsymbol{y}) - \mathbb{E}_F[\rho(\boldsymbol{X}, \boldsymbol{y})] \end{aligned} \tag{4.15}$$

where $\boldsymbol{y} \in \mathbb{R}^d$, $\delta_{\boldsymbol{y}}$ denotes the Dirac mass at y and $\boldsymbol{X}, \boldsymbol{X}'$ are independent random variables following F. Importantly, S_ρ is proper on $\mathcal{P}_\rho$ and, for an ensemble forecast $F = \frac{1}{M} \sum_{m=1}^M \delta_{\boldsymbol{x}_m}$ with M members $\boldsymbol{x}_1, \dots, \boldsymbol{x}_M$, it takes the simple form

$$\mathrm{S}_\rho(F, \boldsymbol{y}) = \frac{1}{2M^2} \sum_{1 \leq m_1, m_2 \leq M} \rho(\boldsymbol{x}_{m_1}, \boldsymbol{x}_{m_2}) + \frac{1}{2}\rho(\boldsymbol{y}, \boldsymbol{y}) - \frac{1}{M} \sum_{m=1}^M \rho(\boldsymbol{x}_m, \boldsymbol{y}), \tag{4.16}$$

making scoring rules particularly useful for ensemble forecasts.

The CRPS is surely the most widely used kernel scoring rule. Equation (4.6) shows that it is associated with the kernel $\rho(x_1, x_2) = |x_1| + |x_2| - |x_1 - x_2|$ (the function $|x_1 - x_2|$ is conditionally semi-definite negative so that ρ is non-negative). For more details on kernel scoring rules, the reader should refer to Gneiting et al. (2005) or Steinwart and Ziegel (2021).

The following proposition reveals that a kernel scoring rule can always be expressed as an aggregation of squared errors (SEs) between transformations of the forecast-observation pair.

Proposition 4.3. *Let S_ρ be the kernel scoring rule associated with the kernel ρ. Then there exists a sequence of transformations $T_l : \mathbb{R}^d \to \mathbb{R}$, $l \geq 1$, such that*

$$\mathrm{S}_\rho(F, \boldsymbol{y}) = \frac{1}{2} \sum_{l \geq 1} \mathrm{SE}(T_l(F), T_l(\boldsymbol{y})),$$

for all predictive distribution $F \in \mathcal{P}_\rho$ and observation $\boldsymbol{y} \in \mathbb{R}^d$.

In particular, the series on the right-hand side is always finite. The proof is provided in Appendix 4.7.3 and relies on the reproducing kernel Hilbert space (RKHS) representation of kernel scoring rules. In particular, we will see that the sequence $(T_l)_{l \geq 1}$ can be chosen as an orthonormal basis of the RKHS associated with the kernel ρ.

This representation of kernel scoring rules can be useful to understand more deeply the comparison of the predictive forecast F and observation $\boldsymbol{y}$. While the definition (4.15) is quite abstract, the series representation can be rewritten

$$\mathrm{S}_\rho(F, \boldsymbol{y}) = \sum_{l \geq 1} \big(\mathbb{E}_F[T_l(\boldsymbol{X})] - T_l(\boldsymbol{y}) \big)^2$$

with X a random variable following F. In other words, for $l \geq 1$, the observed value $T_l(\boldsymbol{y})$ is compared to the predicted value $T_l(\boldsymbol{X})$ under the predictive distribution F using the SE; then all these contributions are aggregated in a series forming the kernel scoring rule.

To give more intuition, we study two important cases in dimension $d = 1$. The details of the computations are provided in Appendix 4.7.3. For the Gaussian kernel scoring rule associated with the kernel

$$\rho(x_1, x_2) = \exp(-(x_1 - x_2)^2/2),$$

some computations yield the series representation

$$\mathrm{S}_\rho(F, y) = \frac{1}{2} \sum_{l \geq 0} \frac{1}{l!} \Big(\mathbb{E}_F[X^l e^{-X^2/2}] - y^l e^{-y^2/2} \Big)^2$$

so that this score compares the probabilistic forecast F and the observation y through the transforms

$$T_l(x) = \frac{1}{\sqrt{l!}} x^l e^{-x^2/2}, \quad l \geq 0.$$

For the CRPS, a possible series representation is obtained thanks to the following wavelet basis of functions: let $T^0(x) = x \mathbb{1}_{[0,1)}(x) + \mathbb{1}_{[1,+\infty)}(x)$ (plateau function) and $T^1(x) = \big(1/2 - |x - 1/2|\big)\mathbb{1}_{[0,1]}(x)$ (triangle function) and consider the collection of functions

$$T^0_l(x) = T^0(x - l), \quad T^1_{l,m}(x) = 2^{-m/2} T^1(2^m x - l), \quad l \in \mathbb{Z}, m \geq 0,$$

where $l \in \mathbb{Z}$ is a position parameter and $m \geq 0$ a scale parameter. Then, the CRPS can be written as

$$\begin{aligned} \mathrm{CRPS}(F, y) &= \sum_{l \in \mathbb{Z}} \mathrm{SE}(T^0_l(F), T^0_l(y)) + \sum_{l \in \mathbb{Z}} \sum_{m \geq 0} \mathrm{SE}(T^1_{l,m}(F), T^1_{l,m}(y)) \\ &= \sum_{l \in \mathbb{Z}} \Big(\mathbb{E}_F[T^0(X - l)] - T^0(y - l) \Big)^2 + \sum_{l \in \mathbb{Z}} \sum_{m \geq 0} 2^{-m} \Big(\mathbb{E}_F[T^1(2^m X - l)] - T(2^m y - l) \Big)^2. \end{aligned}$$

We can see that the CRPS compares forecast and observation through the SE after applying the plateau and triangle transformations for multiple positions and scales and then aggregates all the contributions.

4.4 Applications of the transformation and aggregation principles

4.4.1 Projections

Certainly, the most direct type of transformation is projections of forecasts and observations on their k-dimensional marginals. We denote T_i the projection on the i-th component such that $T_i(\boldsymbol{X}) = X_i$, for all $\boldsymbol{X} \in \mathbb{R}^d$. This allows the forecaster to assess the predictive performance of a forecast for a specific univariate marginal independently of the other variables. If S is a univariate scoring rule proper relative to $\mathcal{P}(\mathbb{R})$, then Proposition 4.1 leads to S_{T_i} being proper relative to $\mathcal{P}(\mathbb{R}^d)$. This "new" scoring rule can be useful if a given marginal is of particular interest (e.g., location of high interest in a spatial forecast). However, it can be more interesting to aggregate such scoring rules across all 1-dimensional marginals. This leads to the following scoring rule

$$\mathrm{S}_{\mathcal{S}_\mathcal{T}, \boldsymbol{w}}(F, \boldsymbol{y}) = \sum_{i=1}^{d} w_i \mathrm{S}_{T_i}(F, \boldsymbol{y}),$$

where $\mathcal{S}_{\mathcal{T}}$ is $\{\mathrm{S}_{T_i}\}_{1\leq i\leq d}$. This setting is popular for assessing the performance of multivariate forecasts and we briefly present examples from the literature falling under this setting. Aggregation of CRPS (4.6) across locations and/or lead times is common practice for plots or comparison tables with uniform weights (Gneiting et al., 2005; Taillardat et al., 2016; Rasp and Lerch, 2018; Schulz and Lerch, 2022b; Lerch and Polsterer, 2022; Hu et al., 2023) or with more complex schemes such as weights proportional to the cosine of the latitude (Ben Bouallègue et al., 2024b). The SE (4.2) and AE (4.3) can be aggregated to obtain RMSE and MAE, respectively (Delle Monache et al., 2013; Gneiting et al., 2005; Lerch and Polsterer, 2022; Pathak et al., 2022). Bremnes (2020) aggregated QSs (4.4) across stations and different quantile levels of interest with uniform weights. Note that the multivariate SE (4.12) can be rewritten as the sum of univariate SE across 1-marginals: $\mathrm{SE}(F, \boldsymbol{y}) = \|\boldsymbol{\mu_F} - \boldsymbol{y}\|_2^2 = \sum_{i=1}^{d} \mathrm{SE}_{T_i}(F, \boldsymbol{y})$.

The second simplest choice is the 2-dimensional case, allowing to focus on pair dependency. We denote $T_{(i,j)}$ the projection on the i-th and j-th components (i.e., the (i, j) pair of components) such that $T_{(i,j)}(\boldsymbol{X}) = X_{i,j} = (X_i, X_j)$. In this setting, S has to be a bivariate proper scoring rule to construct a proper scoring rule $\mathrm{S}_{T_{(i,j)}}$. The aggregation of such scoring rules becomes

$$\mathrm{S}_{\mathcal{S}_{\mathcal{T}},\boldsymbol{w}}(F, \boldsymbol{y}) = \sum_{\substack{i,j=1 \\ i\neq j}}^{d} w_{i,j} \mathrm{S}_{T_{(i,j)}}(F, \boldsymbol{y}).$$

As suggested in Scheuerer and Hamill (2015b) for the VS (4.14), the weights $w_{i,j}$ can be chosen appropriately to optimize the signal-to-noise ratio. For example, in a spatial setting where the dependence between locations is believed to decrease with the distance separating them, the weights $w_{i,j}$ can be chosen to be proportional to the inverse of the distance. This bivariate setting is less used in the literature, we present two articles using or mentioning scoring rules within this scope. In a general multivariate setting, Ziel and Berk (2019) suggests the use of a marginal-copula scoring rule where the copula score is the bivariate copula energy score (i.e., the aggregation of the energy scores across all the regularized pairs). To focus on the verification of the temporal dependence of spatio-temporal forecasts, Ben Bouallègue et al. (2024b) uses the bivariate energy score over consecutive lead times.

In a more general setup, we consider projection on k-dimensional marginals. In order to reduce the number of transformation-based scores to aggregate, it is standard to focus on localized marginals (e.g., belonging to patches of a given spatial size). Denote $\mathcal{P} = \{P_i\}_{1\leq i\leq m}$ a set of valid patches (for some criterion or of a given size) and $\mathcal{S}_{\mathcal{P}}$ the set of transformation-based scores associated with the projections on the patches $\mathcal{P}$. Given a multivariate scoring rule S proper relative to $\mathcal{P}(\mathbb{R}^k)$, we can construct the following aggregated score :

$$\mathrm{S}_{\mathcal{S}_{\mathcal{P}},\boldsymbol{w}}(F, \boldsymbol{y}) = \sum_{P\in\mathcal{P}} w_P \mathrm{S}_P(F, \boldsymbol{y}).$$

This construction can be used to create a scoring rule only considering the dependence of localized components, given that the patches are defined in that sense. The use of patches has similar benefits as the weighting of pairs given a belief on their correlations: obtain a better signal-to-noise ratio and improve the discrimination of the resulting scoring rule. For example, Pacchiardi et al. (2024) introduced patched energy scores as scoring rules to minimize in order to train a generative neural network. The patched energy scores are defined for S = ES and square patches spaced by a given stride. Even though spatial patches may be more intuitive, it is possible to use temporal or spatio-temporal patches. Patch-based scoring rules appear as a natural member of the neighborhood-based methods of the spatial verification classification mentioned in Section 4.2.4. Given that the patches are correctly chosen (e.g., of a size appropriate to the problem at hand), patch-based scoring rules are not subject to the double-penalty

effect.

As noticeable by the low number of examples available in the literature, aggregation (and plain use) of scoring rules based on projection in dimension $k \geq 2$ is not standard practice, probably because such projections may lack interpretability. Instead, to assess the multivariate aspects of a forecast, scoring rules relying on summary statistics are often favored.

4.4.2 Summary statistics

Summary statistics are a central tool of statisticians' toolboxes as they provide interpretable and understandable quantities that can be linked to the behavior of the phenomenon studied. Moreover, their interpretability can be enhanced by the forecaster's experience and this can be leveraged when constructing scoring rules based on them. Summary statistics are commonly present during the verification procedure and this can be extended by the use of new scoring rules derived from any summary statistic of interest. For example, numerous summary statistics can come in handy when studying precipitations over a region covered by gridded observation and forecasts. Firstly, it is common practice to focus on binary events such as the exceedance of a threshold (e.g., the presence or absence of precipitation). This can be studied by using the BS (4.5) on all 1-dimensional marginals as mentioned in the previous subsection but also in a multivariate manner through the fraction of threshold exceedances (FTE) over patches as presented further. Regarding precipitations, it is standard to be interested in the prediction of total precipitation over a region or a time period. This transformation of the field can be leveraged to construct a scoring rule. Finally, it is important to verify that the spatial structure of the forecast matches the spatial structure of observations. The spatial structure can be (partially) summarized by the variogram or by wavelet transformations. The predictive performance for the spatial structure can be assessed by their associated scoring rules: the VS of order p (4.14) and the wavelet-based score (Buschow et al., 2019). Other summary statistics can be of interest to the phenomenon studied, Heinrich-Mertsching et al. (2024) present summary statistics specific to point processes focusing on clustering and intensity.

The most well-known summary statistic is certainly the mean. In spatial statistics, it can be used to avoid double penalization when we are less interested in the exact location of the forecast but rather in a regional prediction. The transformation associated with the mean is

$$\text{mean}_P(\boldsymbol{X}) = \frac{1}{|P|} \sum_{i \in P} X_i, \tag{4.17}$$

where P denotes a patch and $|P|$ its dimension. Proposition 4.1 ensures that this transformation can be used to construct proper scoring rules. The scoring rule involved in the construction has to be univariate, however, the choice depends on the general properties preferred. For example, the SE would focus on the mean of the transformed quantity, whereas the AE would target its median. It is worth noting that the total can be derived by the mean transformation by removing the prefactor

$$\text{total}_P(\boldsymbol{X}) = \sum_{i \in P} X_i.$$

In the case of precipitation, the total is more used than the mean since the total precipitation over a river basin can be decisive in evaluating flood risk. For example, one could construct an adapted version of the amplitude component of the SAL method (Wernli et al., 2008; Radanovics et al., 2018) using the SE if the mean total precipitation is of interest. Gneiting (2011) presents other links between the quantity of interest and the scoring rule associated. Similarly, the transformations associated with the minimum and the maximum over a patch P can be obtained

:

$$\min_P(\boldsymbol{X}) = \min_{i \in P}(X_i);$$
$$\max_P(\boldsymbol{X}) = \max_{i \in P}(X_i).$$

The maximum or minimum can be useful when considering extreme events. It can help understand if the severity of an event is well-captured. For example, as minimum and maximum temperatures affect crop yields (see, e.g., Agnolucci et al. 2020), it can be of particular interest that a weather forecast within an agricultural model correctly predicts the minimum and maximum temperatures. After studying the mean, it is natural to think of the moments of higher order. We can define the transformation associated with the variance over a patch P as

$$\mathrm{Var}_P(\boldsymbol{X}) = \frac{1}{|P|} \sum_{i \in P} (X_i - \mathrm{mean}_P(\boldsymbol{X}))^2.$$

The variance transformation can provide information on the fluctuations over a patch and be used to assess the quality of the local variability of the forecast. In a more general setup, it can be of interest to use a transformation related to the moment of order n and the transformation associated follows naturally

$$\mathrm{M}_{n,P}(\boldsymbol{X}) = \frac{1}{|P|} \sum_{i \in P} X_i^n.$$

More application-oriented transformations are the central or standardized moments (e.g., skewness or kurtosis). Their transformations can be obtained directly from estimators. As underlined in Heinrich-Mertsching et al. (2024), since Proposition 4.1 applies to any transformation, there is no condition on having an unbiased estimator to obtain proper scoring rules.

Threshold exceedance plays an important role in decision making such as weather alerts. For example, MeteoSwiss' heat warning levels are based on the exceedance of daily mean temperature over three consecutive days (Allen et al., 2023a). They can be defined by the simultaneous exceedance of a certain threshold and the fraction of threshold exceedance (FTE) is the summary statistic associated.

$$\mathrm{FTE}_{P,t}(\boldsymbol{X}) = \frac{1}{|P|} \sum_{i \in P} \mathbb{1}_{\{X_i \geq t\}}. \tag{4.18}$$

FTEs can be used as an extension of univariate threshold exceedances and it prevents the double-penalty effect. FTEs may be used to target compound events (e.g., the simultaneous exceedances of a threshold at multiple locations of interest). Roberts and Lean (2008) used an FTE-based SE over different sizes of neighborhoods (patches) to verify at which scale forecasts become skillful. To assess extreme precipitation forecasts, Rivoire et al. (2023) introduces scores for extremes with temporal and spatial aggregation separately. Extreme events are defined as values higher than the seasonal 95% quantile. In the subseasonal-to-seasonal range, the temporal patches are 7-day windows centered on the extreme event and the spatial patches are square boxes of 150 km $\times$ 150 km centered on the extreme event. The final scores are transformed BS (4.5) with a threshold of one event predicted across the patch.

Correctly predicting the structure dependence is crucial in multivariate forecasting. Variograms are summary statistics representing the dependence structure. The variogram of order p of the pair (i, j) corresponds to the following transformation :

$$\gamma_{ij}^p(\boldsymbol{X}) = |X_i - X_j|^p.$$

As mentioned in the Introduction, using both the transformation and aggregation principles, we can recover the VS of order p (4.14) introduced in Scheuerer and Hamill (2015b) :

$$\mathrm{VS}_p(F, \boldsymbol{y}) = \sum_{i,j=1}^{d} w_{ij} \mathrm{SE}_{\gamma_{ij}^p}(F, \boldsymbol{y}) = \sum_{i,j=1}^{d} w_{ij} \left(\mathbb{E}_F[|X_i - X_j|^p] - |y_i - y_j|\right)^2 .$$

Along with the well-known VS of order p, Scheuerer and Hamill (2015b) introduced alternatives where the scoring rule applied on the transformation is the CRPS (4.6) or the AE (4.3) instead of the SE (4.2). As mentioned previously, under the *intrinsic hypothesis* of Matheron (1963) (i.e., pairwise differences only depend on the distance between locations), the weights can be selected to obtain an optimal signal-to-noise ratio. Moreover, the weights could be selected to investigate a specific scale by giving a non-zero weight to pairs separated by a given distance.

In the case of spatial forecasts over a grid of size $d \times d$, a spatial version of the variogram transformation is available :

$$\gamma_{\boldsymbol{i},\boldsymbol{j}}(\boldsymbol{X}) = |X_{\boldsymbol{i}} - X_{\boldsymbol{j}}|^p,$$

where $\boldsymbol{i}, \boldsymbol{j} \in \mathcal{D} = \{1, \ldots, d\}^2$ are the coordinates of grid points. Under the intrinsic hypothesis of Matheron (1963), the variogram between grid points separated by the vector $\boldsymbol{h}$ can be estimated by :

$$\gamma_{\boldsymbol{X}}(\boldsymbol{h}) = \frac{1}{2|\mathcal{D}(\boldsymbol{h})|} \sum_{\boldsymbol{i} \in \mathcal{D}(\boldsymbol{h})} \gamma_{\boldsymbol{i},\boldsymbol{i}+\boldsymbol{h}}(\boldsymbol{X}),$$

where $\mathcal{D}(\boldsymbol{h}) = \{\boldsymbol{i} \in \mathcal{D} : \boldsymbol{i}+\boldsymbol{h} \in \mathcal{D}\}$. This directed variogram can be used to target the verification of the anisotropy of the dependence structure. The isotropy transformation associated to the distance h can be defined by

$$T_{\mathrm{iso},h}(\boldsymbol{X}) = -\frac{\left(\gamma_X((h,0)) - \gamma_X((0,h))\right)^2}{\frac{2\gamma_X((h,0))^2}{|\mathcal{D}((h,0))|} + \frac{2\gamma_X((0,h))^2}{|\mathcal{D}((0,h))|}}. \tag{4.19}$$

This transformation is the isotropy pre-rank function proposed in Allen et al. (2024). The isotropy transformation considers the orthogonal directions formed by the abscissa and ordinate axes and evaluates how the variogram changes between these directions. The transformation leads to negative or zero quantities with values close to zero characterizing isotropy and negative values corresponding to the anisotropy of the variograms in the directions and at the scale involved.

4.4.3 Other transformations

Transformations other than projections or summary statistics can be used to target forecast characteristics. For example, a transformation in the form of a change of coordinates or a change of scale (e.g., a logarithmic scale) can be used to obtain proper scoring rules. We highlight two families of scoring rules that can be seen as transformation-based scoring rules: wavelet-based scoring rules and threshold-weighted scoring rules.

Generally speaking, wavelet-based scoring rules are built thanks to a projection of forecast and observation fields onto a wavelet basis. Based on the wavelet coefficients, dimension reduction might be performed to target specific characteristics such as the dependence structure or the location. The resulting coefficients of the forecast fields are compared to the coefficients of the observations fields using scoring rules (e.g., squared error (SE) or energy score (ES)). Wavelet transformations are (complex) transformations, and thus, the scoring rules associated fall within the scope of Proposition 4.1. In particular, Buschow et al. (2019) used a dimension reduction procedure resulting in the obtention of a mean and a scale spectra and used scoring

rules to compare forecasts and observation spectra. For example, the ES of the mean spectrum is used and shows good discrimination ability when the scale structure is misspecified.

Note that Buschow et al. (2019) proposed two other wavelet-based scoring rules: one based on the earth mover's distance (EMD) of the scale histograms and one based on the distance in the scale histograms' center of mass. The EMD-based scoring rules are not proper since the EMD is not a proper scoring rule (Thorarinsdottir et al., 2013) and the so-called distance between centers of mass is not a distance but rather a difference of position leading to an improper scoring rule. However, the ES-based scoring rules are proper and could be derived from scale histograms. Despite their apparent complexity, wavelet transformations allow to target interpretable characteristics such as the location (Buschow, 2022), the scale structure (Buschow et al., 2019; Buschow and Friederichs, 2020) or the anisotropy (Buschow and Friederichs, 2021). The transformations proposed for the deterministic forecasts setting in most of these articles could be used as foundations for future work willing to propose wavelet-based proper scoring rules targeting the location, the scale structure or the anisotropy.

As showcased in Heinrich-Mertsching et al. (2024) for a specific example and hinted in Allen et al. (2024), transformations can also be used to emphasize certain outputs. Threshold weighting is one of the three main types of weighting conserving the propriety of scoring rules. Its name comes from the fact that it corresponds to a weighting over different thresholds in the case of CRPS (4.7) (Gneiting, 2011). Recall that given a conditionally negative definite kernel ρ, the kernel scoring associated S_ρ (4.15) is proper relative to $\mathcal{P}_\rho$. Many popular scoring rules are kernel scores such as the BS (4.5), the CRPS (4.6), the ES (4.13) and the VS (4.14). By definition (Allen et al., 2023b, Definition 4), threshold-weighted kernel scores are constructed as

$$\begin{aligned}\mathrm{twS}_\rho(F, \boldsymbol{y}; v) &= \mathbb{E}_F[\rho(v(\boldsymbol{X}), v(\boldsymbol{y}))] - \frac{1}{2}\mathbb{E}_F[\rho(v(\boldsymbol{X}), v(\boldsymbol{X}'))] - \frac{1}{2}\rho(v(\boldsymbol{y}), v(\boldsymbol{y}));\\ &= \mathrm{S}_\rho(v(F), v(\boldsymbol{y})),\end{aligned}$$

where v is the chaining function capturing how the emphasis is put on certain outputs. With this explicit definition, it is obvious that threshold-weighted kernel scores are covered by the framework of Proposition 4.1. It can be noted that Proposition 4 in Allen et al. (2023b) states that strict propriety of the kernel scoring rule is preserved by the chaining function v if and only if v is injective. Weighted scoring rules allow to emphasize particular outcomes: when studying extreme events, it is often of particular interest to focus on values larger than a given threshold t and this can be achieved using the chaining rule $v(x) = \mathbb{1}_{x \geq t}$. Threshold-weighted scoring rules have been used in verification procedures in the literature; we illustrate its use through three different studies. Lerch and Thorarinsdottir (2013) aggregated across stations twCRPS to compare the upper tail performance of different daily maximum wind speed forecasts. Chapman et al. (2022) aggregated the threshold-weighted CRPS across locations to study the improvement of statistical postprocessing techniques, the importance of predictors and the influence of the size of the training set on the performance. Allen et al. (2023a) used threshold-weighted versions of the CRPS, the ES, and the VS to compare the predictive performance of forecasts regarding heatwave severity; the scoring rules were aggregated across stations. Readers may refer to Allen et al. (2023a) and Allen et al. (2023b) for insightful reviews of weighted scoring rules in both univariate and multivariate settings.

4.5 Simulation study

This section provides simulated examples to showcase the different uses of the framework introduced in Section 4.3 to construct interpretable proper scoring rules for multivariate forecasts.

Four examples are developed. Firstly, a setup where the emphasis is put on 1-marginal verification is proposed. This setup serves as a means of recalling and showing the limitations of strictly proper scoring rules and the benefits of interpretable scoring rules in a concrete setting. Secondly, a standard multivariate setup is studied where popular multivariate scoring rules (i.e., VS and ES) are compared to a multivariate scoring rule aggregated over patches and an aggregation-and-transformation-based scoring rule in their discrimination ability regarding the dependence structure. Thirdly, a setup introducing anisotropy in both observations and forecasts is introduced. The anisotropic score is constructed based on the transformation principle with the goal of discriminating differences of anisotropy in the dependence structure between forecast and observations. Fourthly, we propose a setup to test the sensitivity of scoring rules to the double-penalty effect and we introduce scoring rules that can be built to be resilient to some manifestation of the double-penalty effect.

In these four numerical experiments, the spatial field is observed and predicted on a regular 20×20 grid $\mathcal{D} = \{1, \dots, 20\} \times \{1, \dots, 20\}$. Observations are realizations of a Gaussian random field $(G(s))_{s \in \mathcal{D}}$ with zero mean and power-exponential covariance defined as

$$\operatorname{cov}(G(s), G(s')) = {\sigma_0}^2 \exp\left(- \left(\frac{\|s - s'\|}{\lambda_0} \right)^{\beta_0} \right), \quad s, s' \in \mathcal{D}. \tag{4.20}$$

The parameters are taken equal to $\sigma_0 = 1$, $\lambda_0 = 3$ and $\beta_0 = 1$.

In each numerical experiment, we compare a few predictive distributions, including the distribution generating observations and other ones deviating from the generative distributions in a specific way. These different predictive distributions are evaluated with different scoring rules and the aim is to illustrate the discriminatory ability of the different scoring rules.

The simulation study uses 500 observations of the random field $(G(s))_{s \in \mathcal{D}}$. The scoring rules are computed using exact formulas when possible (see Appendix 4.7.5), and, when exact formulas are not available, they are computed based on a sample of size 100 (i.e., ensemble forecasts) of the probabilistic forecast. Estimated expectations over the 500 observations are computed and the experiment is repeated 10 times. The corresponding results are represented by boxplots. The units of the scoring rules are rescaled by the average expected score of the true distribution (i.e., the ideal forecast). The statistical significativity of the ranking between forecasts is tested using a Diebold-Mariano test (Diebold and Mariano, 1995). When deemed necessary, statistical significativity is mentioned for a confidence level of 95%.

The code used for the different numerical experiments is publicly available[4].

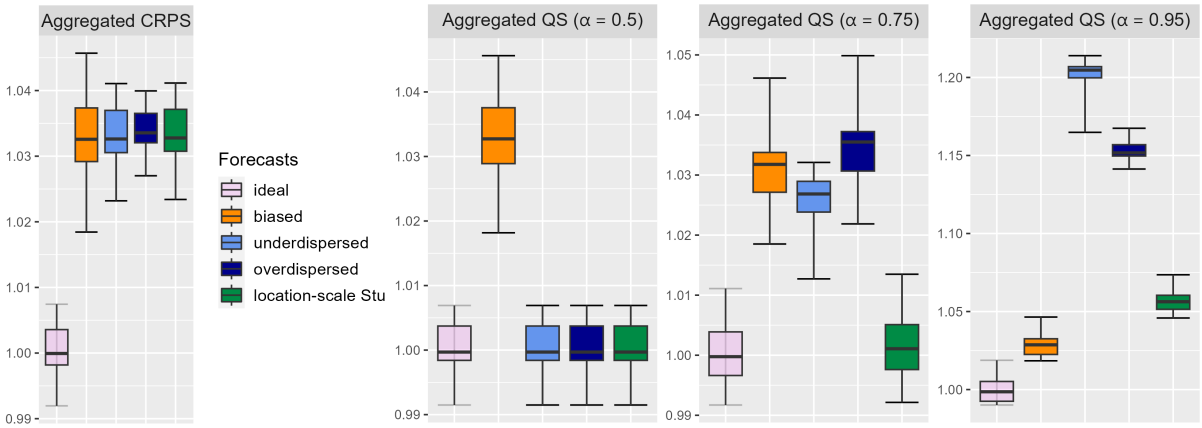
4.5.1 Marginals

This first numerical experiment focuses on the prediction of the 1-dimensional marginal distributions and the aggregation of univariate scoring rules. For simplicity, we consider only stationary random fields so that the 1-marginal distribution is the same at all grid points. Although similar conclusions could be drawn from a univariate framework (i.e., with independent 1-dimensional rather than spatial observations), this example aims to clarify the notion of interpretability and presents notions that will be reused in the following examples. The verification of marginals, along with other simple quantities, is usually one of the first steps of any multivariate forecast verification process.

Observations follow the model of (4.20) and multiple competing forecasts are considered:

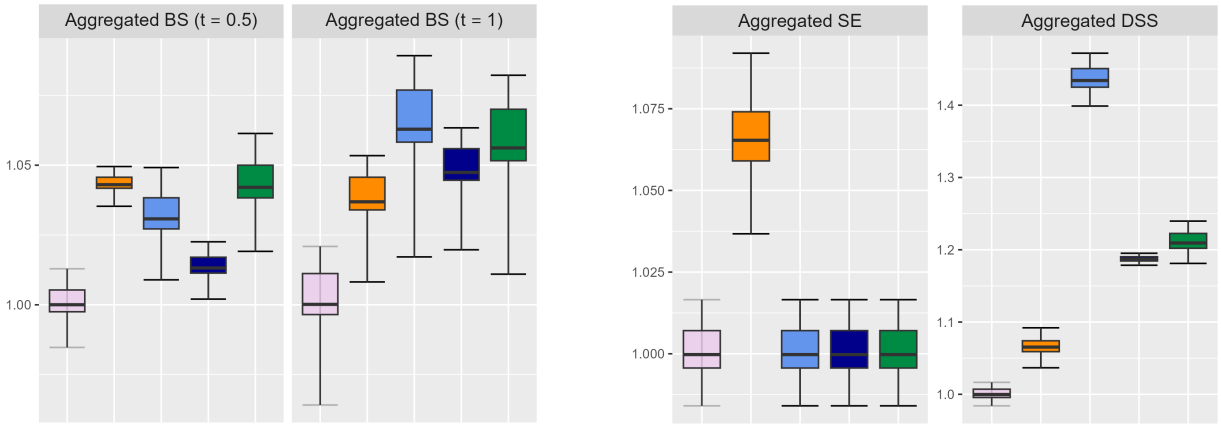
- the *ideal forecast* is the Gaussian distribution generating observations and is used as a reference;
- the *biased forecast* is a Gaussian predictive distribution with the same covariance structure as the observation but a different mean $\mathbb{E}[F_{\text{bias}}(s)] = c = 0.255$;

[4] https://github.com/pic-romain/aggregation-transformation



(a) Aggregated CRPS

(b) Aggregated QS



(c) Aggregated BS

(d) Aggregated DSS and SE

Figure 4.1: Expectation of aggregated univariate scoring rules: (a) the CRPS, (b) the quantile score, (c) the Brier score, and (d) the squared error and the Dawid-Sebastiani score, for the ideal forecast (light violet), a biased forecast (orange), an under-dispersed forecast (lighter blue), an over-dispersed forecast (darker blue) and a local-scale Student forecast (green). More details are available in the main text.

- the *overdispersed forecast* and the *underdispersed forecast* are Gaussian predictive distributions from the same model as the observations except for an overestimation ($\sigma = 1.4$) and an underestimation ($\sigma = 2/3$) of the variance respectively;
- the *location-scale Student forecast* is used where the marginals follow location-scale Student-t distributions with parameters $\mu = 0$, $df = 5$, and τ is such that the standard deviation is 0.745 and the covariance structure the same as in (4.20).

In order to compare the predictive performance of forecasts, we use scoring rules constructed by aggregating univariate scoring rules. Here, the aggregation is done with uniform weights since there is no prior knowledge on the locations. The univariate scoring rules considered are the continuous ranked probability score (CRPS), the Brier score (BS), the quantile score (QS), the squared error (SE) and the Dawid-Sebastiani score (DSS). Figure 4.1a compares five different forecasts based on their expected CRPS. It can be seen that all forecasts except for the ideal one have similar expected values and no sub-efficient forecast is significantly better than the others. In order to gain more insight into the predictive performance of the forecast, it is necessary to use other scoring rules. In practice, the distribution is unknown; thus, it is impossible to know if a

forecast is efficient; it is only possible to provide a ranking linked to the *closeness* of the forecast with respect to the observations. The definition of closeness depends on the scoring rule used: for example, the CRPS defines closeness in terms of the integrated quadratic distance between the two cumulative distribution functions (see, e.g., Thorarinsdottir and Schuhen 2018).

If the quantity of interest is the value of a quantile of a certain level α, the aggregated QS is an appropriate scoring rule. Figure 4.1b shows the expected aggregated QS for three different levels α : $\alpha = 0.5$, $\alpha = .75$ and $\alpha = 0.95$. $\alpha = 0.5$ is associated with the prediction of the median and, since all the forecasts are symmetric and only the biased forecast is not centered on zero, the other forecasts are equally the best and efficient forecasts. If the third quartile is of interest ($\alpha = 0.75$), the location-scale Student forecast appears as significantly the best (among the non-ideal). For the higher level of $\alpha = 0.95$, the biased forecast is significantly the best since its bias error seems to be compensated by its correct prediction of the variance. Depending on the level of interest, the best forecast varies; the only forecast that would appear to be the best regardless of the level α is the ideal forecast, as implied by (4.8).

If a quantity of interest is the exceedance of a threshold t at each location, then the aggregated BS is an interesting scoring rule. Figure 4.1c shows the expectation of aggregated BS for the different forecasts and for two different thresholds ($t = 0.5$ and $t = 1$). Among the non-ideal forecasts, there seems to be a clearer ranking than for the CRPS. The overdispersed forecast is significantly the best regarding the prediction of the exceedance of the threshold $t = 0.5$ and the biased forecast is significantly the best regarding the exceedance of $t = 1$. As for the aggregated quantile score, the best forecast depends on the threshold t considered and the only forecast that is the best regardless of the threshold t is the ideal one (see Eq. (4.7)).

If the moments are of interest, the aggregated SE discriminates the first moment (i.e., the mean) and the aggregated DSS discriminates the first two moments (i.e., the mean and the variance). Figure 4.1d presents the expected values of these scoring rules for the different forecasts considered in this example. The aggregated SEs of all forecasts, except the biased forecast, are equal since they have the same (correct) marginal means. The aggregated DSS presents the biased forecast as significantly the best one (among non-ideal). This is caused by the combined discrimination of the first two moments of the Dawid-Sebastiani score (see Eq. (4.9) and Appendix 4.7.1).

4.5.2 Multivariate scores over patches

This second numerical experiment focuses on the prediction of the dependence structure. Observations are sampled from the model of Eq. (4.20) and we compare forecasts that differ only in their dependence structure through misspecification of the range parameter λ and the smoothness parameter β:

- the *ideal forecast* is the Gaussian distribution generating the observations;
- the *small-range forecast* and the *large-range forecast* are Gaussian predictive distributions from the same model (4.20) as the observations except for an underestimation ($\lambda = 1$) and an overestimation ($\lambda = 5$), respectively, of the range;
- the *under-smooth forecast* and the *over-smooth forecast* are Gaussian predictive distributions from the same model as the observations except for an underestimation ($\beta = 0.5$) and an overestimation ($\beta = 2$), respectively, of the smoothness.

Since the forecasts differ only in their dependence structure, scoring rules acting on the 1-dimensional marginals would not be able to distinguish the ideal forecast from the others. We use the variogram score (VS) as a reference since it is known to discriminate misspecification of the dependence structure. We introduce the patched energy score, which results from the

aggregation of the ES (with $\alpha = 1$) over patches, defined as

$$\mathrm{ES}_{\mathcal{P},\boldsymbol{w}_\mathcal{P}}(F, \boldsymbol{y}) = \sum_{P \in \mathcal{P}} w_P \mathrm{ES}_1(F_P, \boldsymbol{y}_P),$$

where $\mathcal{P}$ is an ensemble of spatial patches, w_P is the weight associated with a patch $P \in \mathcal{P}$ and F_P is the marginal of F over the patch P. In order to make the scoring more interpretable, only square patches of a given size s are considered and the weights w_P are uniform ($w_P = 1/|\mathcal{P}|$). Moreover, we consider the aggregated CRPS and the ES since they are limiting cases of the patched ES for 1×1 patches and a single patch over the whole domain $\mathcal{D}$, respectively. Additionally, we proposed the p-variation score (pVS), which is based on the p-variation transformation to focus on the discrimination of the regularity of the random fields,

$$T_{p-var,\boldsymbol{s}}(\boldsymbol{X}) = |\boldsymbol{X}_{\boldsymbol{s}+(1,1)} - \boldsymbol{X}_{\boldsymbol{s}+(1,0)} - \boldsymbol{X}_{\boldsymbol{s}+(0,1)} + \boldsymbol{X}_{\boldsymbol{s}}|^p$$

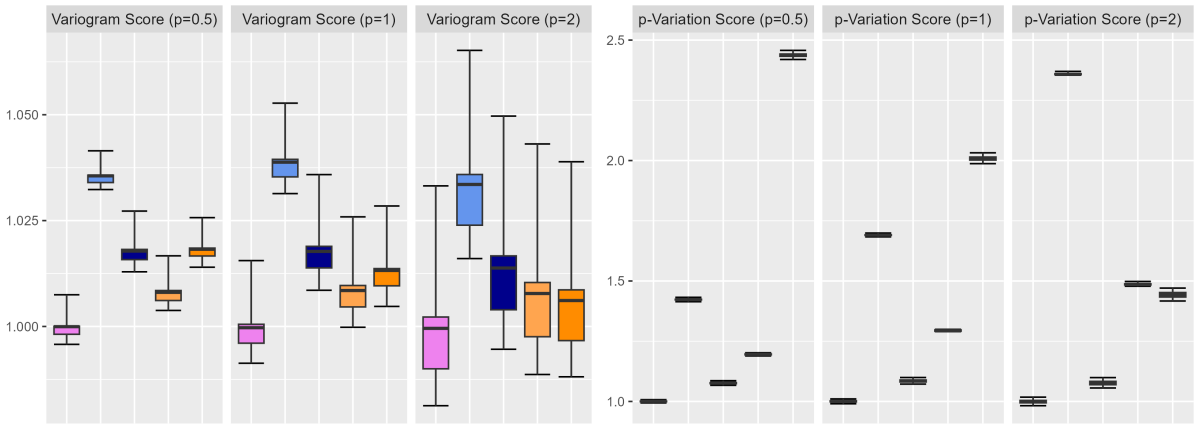
$$\begin{aligned} p\mathrm{VS}(F, \boldsymbol{y}) &= \sum_{\boldsymbol{s} \in \mathcal{D}^*} w_{\boldsymbol{s}} \mathrm{SE}_{T_{p-var,\boldsymbol{s}}}(F, \boldsymbol{y}); \\ &= \sum_{\boldsymbol{s} \in \mathcal{D}^*} w_{\boldsymbol{s}} (\mathbb{E}_F[T_{p-var,\boldsymbol{s}}(\boldsymbol{X})] - T_{p-var,\boldsymbol{s}}(\boldsymbol{y}))^2, \end{aligned}$$

where $\mathcal{D}^*$ is the domain $\mathcal{D}$ restricted to grid points such that $T_{p-var,\boldsymbol{s}}$ is defined (i.e., $\mathcal{D}^* = \{1, \dots, 19\} \times \{1, \dots, 19\}$). Note that in the literature on fractional random fields, the p-variation is an important characteristic used to characterize the roughness of a random field and is commonly used for estimation purposes, see Benassi et al. (2004), Basse-O'Connor et al. (2021) and the references therein.

In Figure 4.2, the ES and the patched ES were computed using samples from the forecasts since closed expressions could not be derived. However, closed formulas for the VS and the pVS were derived and are available in Appendix 4.7.5. As already shown in Scheuerer and Hamill (2015b), the VS is able to significantly discriminate misspecification of the dependence structure induced by the range parameter λ (see Fig. 4.2a). Smaller orders of p (such as $p = 0.5$) appear as more informative than higher ones. Moreover, it is able to discriminate misspecification induced by the smoothness parameter β (significantly for all orders p considered) even if it is less marked than for the misspecification of the range λ.

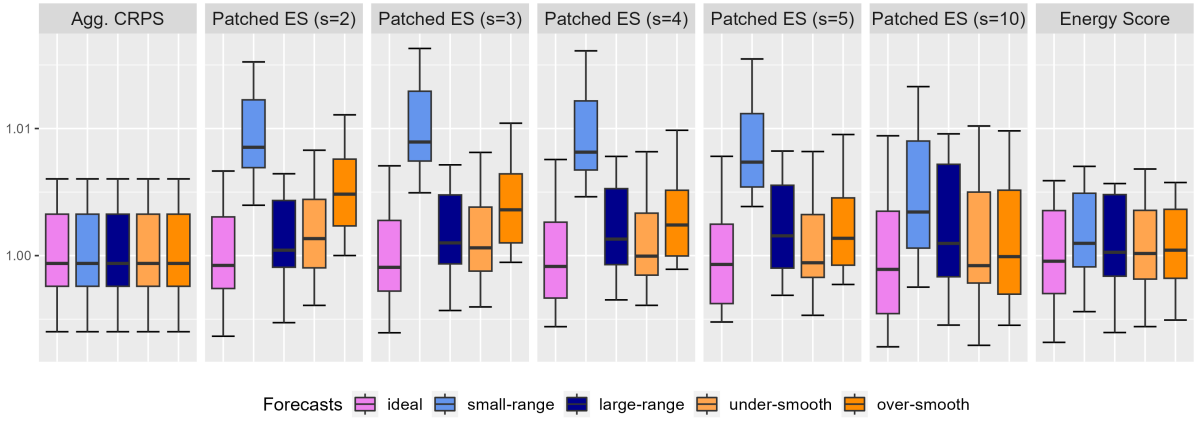
Figure 4.2b compares the forecasts using the p-variation score with $p \in \{0.5, 1, 2\}$. Note that the forecasts are provided in the same order as in the other sub-figures. The pVS is able to (significantly) discriminate all four sub-efficient forecasts from the ideal forecast at all order p. In the cases considered, the pVS has a stronger discriminating ability than the VS; in particular, for misspecification of the smoothness parameter β. The overall improvement in the discrimination ability of the pVS compared to the VS is due to the fact that it only considers local pair interactions between grid points; which in the experimental setup considered greatly improves the signal-to-noise ratio compared to the VS. For example, it would be incapable of differentiating two forecasts that only differ in their longer-range dependence structure, where the VS should discriminate the two forecasts.

Figure 4.2c shows that the patched ESs have a better discrimination ability than the ES. As expected by the clear analogy between the variogram score weights and the selection of valid patches, focusing on smaller patches improves the signal-to-noise ratio. For all patch size s considered, the patched ES significantly discriminates the ideal forecast from the others. Whereas the ES does not significantly discriminate the misspecification of smoothness of the under-smooth and over-smooth forecasts. Nonetheless, the patched ES remains less sensitive than the VS to misspecifications in the dependence structure through the range parameter λ or



(a) Variogram score

(b) p-Variation score



(c) Aggregated CRPS, patched ESs and ES

Figure 4.2: Expectation of scoring rules focused the dependence structure: (a) the variogram score, (b) the p-variation score and (c) the patched energy score (and its limiting cases: the aggregated CRPS and the energy score), for the ideal forecast (violet), the small-range forecast (lighter blue), the large-range forecast (darker blue), the under-smooth forecast (lighter orange), and the over-smooth forecast (darker orange). More details are available in the main text.

the smoothness parameter β.

The VS relies on the aggregation and transformation principles and is able to discriminate the dependence structure. Similarly, the pVS is able to discriminate misspecifications of the dependence structure. Being based on more local transformations (i.e., p-variation transformation instead of variogram transformation), it has a greater discrimination ability than the VS in this experimental setup. In addition to this known application of the aggregation and transformation principles, it has been shown that multivariate transformations can be used to obtain patched scores that, in the case of the ES, lead to an improvement in the signal-to-noise ratio with respect to the original scoring rule.

4.5.3 Anisotropy

In this example, we focus on the anisotropy of the dependence structure. We introduce geometric anisotropy in observations and forecasts via the covariance function in the following way

$$\operatorname{cov}(G(s), G(s')) = \exp\left(-\left(\frac{\|s - s'\|_A}{\lambda_0}\right)\right)$$

with $\|s - s'\|_A = (s - s')^T A(s - s')$. The matrix A has the following form :

$$A = \begin{bmatrix} \cos\theta & -\sin\theta \\ \rho\sin\theta & \rho\cos\theta \end{bmatrix}$$

with $\theta \in [-\pi/2, \pi/2]$ the direction of the anisotropy and ρ the ratio between the axes.

The observations follow the anisotropic version of the model in Eq. (4.20) where the covariance function presents the geometric anisotropy introduced above with $\lambda_0 = 3$ (as previously) and $\rho_0 = 2$ and $\theta_0 = \pi/4$. Multiple forecasts are considered that only differ in their prediction of the anisotropy in the model:

- the *ideal forecast* has the same distribution as the observations and is used as a reference;
- the *small-angle forecast* and the *large-angle forecast* have a correct ratio ρ but an under- and over-estimation of the angle, respectively (i.e., $\theta_{\text{small}} = 0$ and $\theta_{\text{large}} = \pi/2$);
- the *isotropic forecast* and the *over-anisotropic forecast* have a ratio $\rho = 1$ and $\rho = 3$, respectively, but a correct angle θ.

Since these forecasts differ only in the anisotropy of their dependence structure, scoring rules not suited to discriminate the dependence structure would not be able to differentiate them. We compare two proper scoring rules: the variogram score and the anisotropic scoring rule. The variogram score is studied in two different settings: one where the weights are proportional to the inverse of the distance and one where the weights are proportional to the inverse of the anisotropic distance $\|\cdot\|_A$, which is supposed to be more informed since it is the quantity present in the covariance function. The anisotropic score (AS) is a scoring rule based on the framework introduced in Section 4.3 and, in its general form, it is defined as

$$\text{AS}(F, \boldsymbol{y}) = \sum_h w_h \text{S}_{T_{\text{iso},h}}(F, \boldsymbol{y}) = \sum_h w_h \text{S}(T_{\text{iso},h}(F), T_{\text{iso},h}(\boldsymbol{y})), \tag{4.21}$$

where $T_{\text{iso},h}$ is a transformation summarizing the anisotropy of a field such as the one introduced in (4.19). Additionally, we use a special case of this scoring rule where we do not aggregate across the scales h and where S is the squared error :

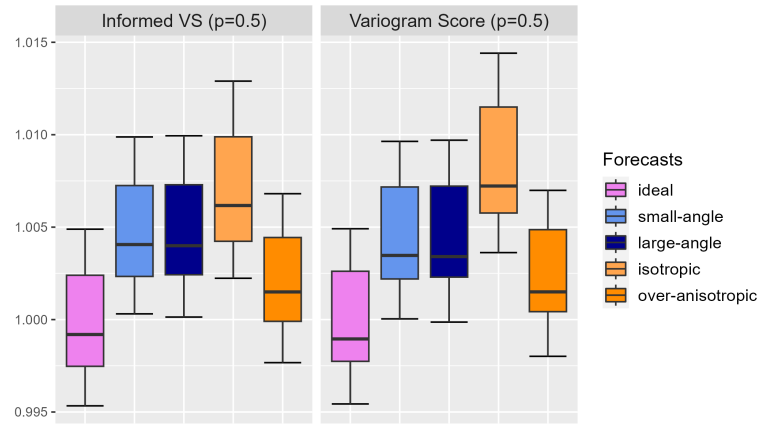
$$S_{T_{\text{iso},h}}(F, \boldsymbol{y}) = \text{SE}(T_{\text{iso},h}(F), T_{\text{iso},h}(\boldsymbol{y})) = \left(\mathbb{E}_{T_{\text{iso},h}(F)}[X] - T_{\text{iso},h}(\boldsymbol{y})\right)^2. \tag{4.22}$$

We use a transformation similar to the one of (4.19) where instead the axes are the first and second bisectors. This leads to the following formula:

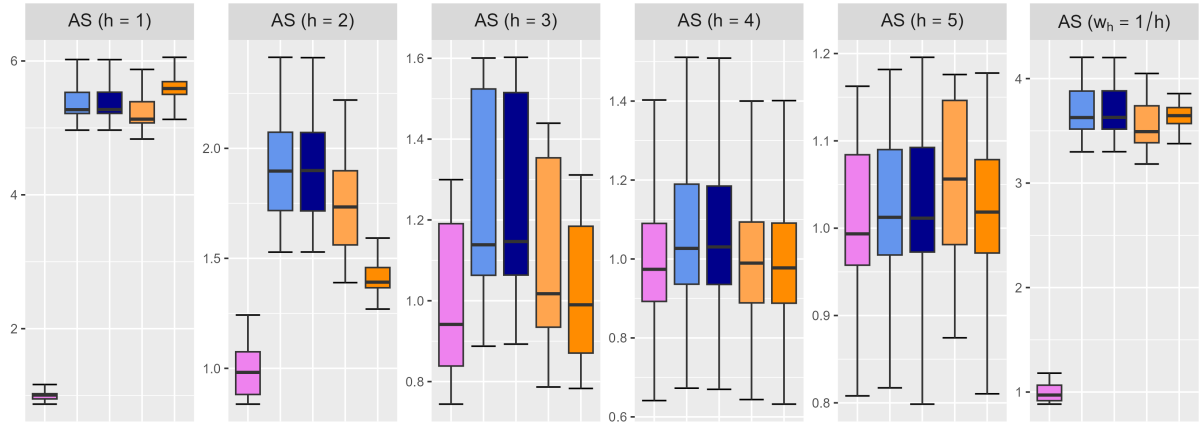
$$T_{\text{iso},h}(\boldsymbol{X}) = -\frac{\left(\gamma_X((h,h)) - \gamma_X((-h,h))\right)^2}{\frac{2\gamma_X((h,h))^2}{|\mathcal{D}((h,h))|} + \frac{2\gamma_X((-h,h))^2}{|\mathcal{D}((-h,h))|}}.$$

The choice of this transformation instead of the transformation based on the anisotropy along the abscissa and ordinate is motivated by the fact that it leads to a clearer differentiation of the forecasts (not shown).

Figure 4.3a presents the variogram score of order $p = 0.5$ in its two variants. Both the standard VS and the informed VS are able to significantly discriminate the ideal forecast from the other forecasts but they have a weak sensitivity to misspecification of the geometric anisotropy. Even though the informed VS is supposed to increase the signal-to-noise ratio compared to the standard VS; it is not more sensitive to misspecifications in the experimental setup considered.



(a) Variogram score



(b) Anisotropic score for different scales h and aggregated across scales ($w_h = 1/h$)

Figure 4.3: Expectation of interpretable proper scoring rules focused the dependence structure: (a) the variogram score and (b) the anisotropic score, for the ideal forecast (violet), the small-angle forecast (lighter blue), the large-angle forecast (darker blue), the isotropic forecast (lighter orange) and the over-anisotropic forecast (darker orange). More details are available in the main text.

Other orders of variograms were tested and worsened the discrimination ability of both standard and informed VS (not shown).

Figure 4.3b shows the AS (4.22) with scales $1 \leq h \leq 5$ for the different forecasts and the aggregated AS (4.21), where the scales are aggregated with weights $w_h = 1/h$. The anisotropic scores were computed using samples drawn from the forecasts; this explains why the ideal forecast may appear sub-efficient for some values of h (e.g., $h = 4$). As aimed by its construction, the AS is able to significantly distinguish the correct anisotropy behavior in the dependence structure for values of h up to $h = 3$ included. For $h = 4$, the AS does not significantly discriminate the isotropic forecast and the over-anisotropic forecast from the ideal one. The fact that $h = 1$ is the most sensitive to misspecifications is probably caused by the fact that the strength of the dependence structure decays with the distance (i.e., with h). This shows that the hyperparameter h plays an important role in having an informative AS (as do the weights and the order p for the variogram score). For $h = 2$ in particular, it can be seen that the AS is not sensitive to the misspecification of the ratio ρ and the angle θ in the same manner. This depends on the degree of misspecification but also on the hyperparameters of the AS. The aggregated AS allows us to avoid the selection of a scale h while maintaining the discrimination ability of the lower values of h.

The anisotropic score is an interpretable scoring rule targeting the anisotropy of the dependence structure. However, it has the limitation of introducing hyperparameters in the form of the scale h and the axes along which the anisotropy is measured. Aggregation across scales and axes can circumvent the selection of these hyperparameters; however, a clever choice of weights will be required to maintain the signal-to-noise ratio.

4.5.4 Double-penalty effect

In this example, we illustrate in a simple setting how scoring rules over patches can be robust to the double-penalty effect (see Section 4.2.4). The double-penalty effect is introduced in the form of forecasts that deviate from the ideal forecast by an additive or multiplicative noise term (i.e., nugget effect). The noises are centered uniforms such that the forecasts are correct on average but incorrect over each grid point.

Observations follow the model of (4.20) with the parameters $\sigma_0 = 1$, $\lambda_0 = 3$ and $\beta_0 = 1$. As per usual the *ideal forecast*, having the same distribution as the observations, is used as a reference. *Additive-noised forecasts* are the first type of forecast introduced to test the sensitivity of scoring rules to the form of the double-penalty effect (presented above). They differ from the ideal forecast through their marginals in the following way :

$$F_{\text{add}}(s) = \mathcal{N}(\epsilon_s, \sigma_0^2),$$

where $\epsilon_s \sim \text{Unif}([-r, r])$ is a spatial white noise independent at each location $s \in \mathcal{D}$. This has an effect on the mean of the marginals at each grid point. Three different noise range values are tested $r \in \{0.1, 0.25, 0.5\}$. Similarly, *multiplicative-noised forecasts* that differ from the ideal forecast through their marginals are introduced :

$$F_{\text{mul}}(s) = \mathcal{N}(0, \sigma^2(1 + \eta_s)^2),$$

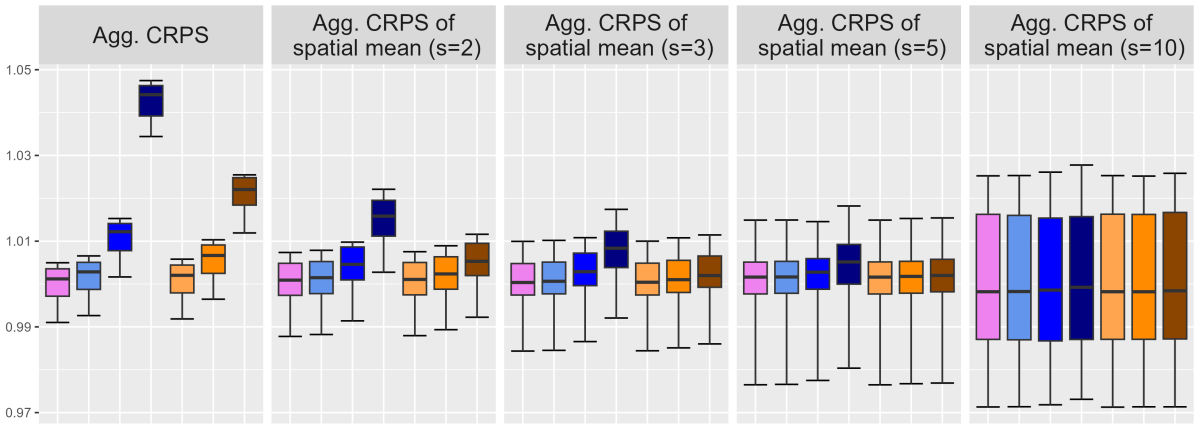
where $\eta_s \sim \text{Unif}([-r, r])$ and $s \in \mathcal{D}$. This has an effect on the variance of the marginals at each grid point and, thus, on the covariance. The same noise range values are tested $r \in \{0.1, 0.25, 0.5\}$.

The aggregated CRPS is a naive scoring rule that is sensitive to the double-penalty effect. We propose the aggregated CRPS of spatial mean which is defined as

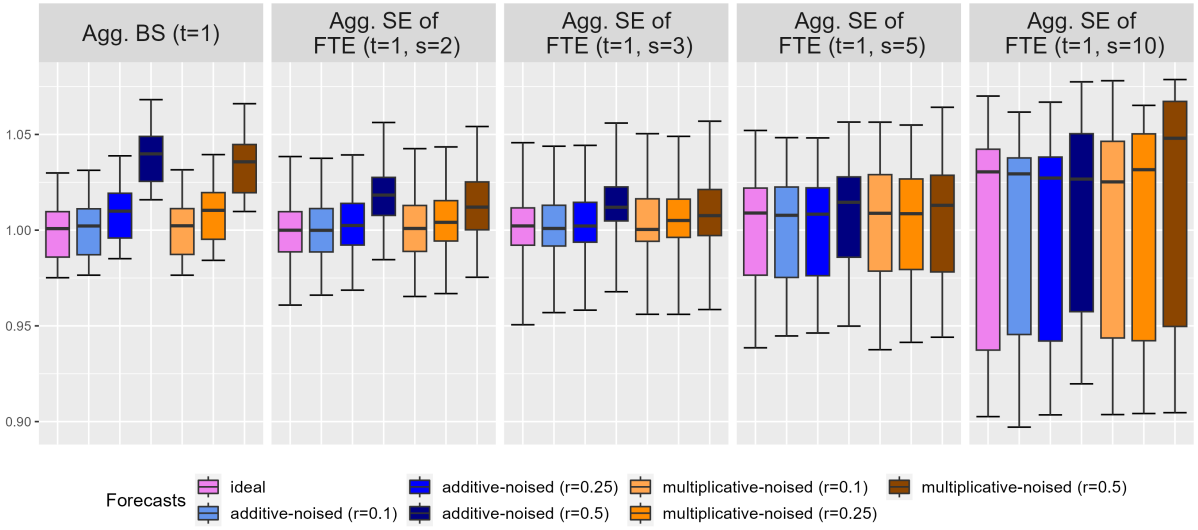
$$\begin{aligned}\text{CRPS}_{\text{mean}_{\mathcal{P}}, \boldsymbol{w_{\mathcal{P}}}}(F, \boldsymbol{y}) &= \sum_{P \in \mathcal{P}} w_P \text{CRPS}_{\text{mean}_P}(F, \boldsymbol{y}); \\ &= \sum_{P \in \mathcal{P}} w_P \text{CRPS}(\text{mean}_P(F), \text{mean}_P(\boldsymbol{y})),\end{aligned}$$

where $\mathcal{P}$ is an ensemble of spatial patches, w_P is the weight associated with a patch $P \in \mathcal{P}$ and mean_P the spatial mean over the patch P (4.17). It is a proper scoring rule, and it has an interpretation similar to the aggregated CRPS, but the forecasts are only evaluated on the performance of their spatial mean. In order to make the scoring more interpretable, only square patches of a given size s are considered and the weights w_P are uniform. The size of the patches s can be determined by multiple factors such as the physics of the problem, the constraints of the verification in the case of models on different scales, or hypotheses on a different behavior below and above the scale of the patch (e.g., independent and identically distributed; Taillardat and Mestre 2020). Note that the aggregated CRPS of spatial mean is equal to the aggregated CRPS when patches of size $s = 1$ are considered.

If a quantity of interest is the exceedance of a threshold t, the scoring rule associated with that is the Brier score (4.5). We compare the aggregated BS with its counterpart over patches:



(a) Aggregated CRPS and CRPS of spatial mean



(b) Aggregated BS and SE of FTE

Figure 4.4: Expectation of scoring rules tested on their sensitivity to double-penalty effect : (a) the aggregated CRPS and the aggregated CRPS of spatial mean, and (b) the aggregated Brier score and the aggregated squared error of fraction of threshold exceedances, for the ideal forecast (violet), the additive-noised forecasts (shades of blue), and the multiplicative-noised forecasts (shades of orange). For the noised forecasts, darker colors correspond to larger values of the range $r \in \{0.1,\ 0.25,\ 0.5\}$. More details are available in the main text.

the aggregated SE of the FTE. It is defined as

$$\begin{aligned}
\mathrm{SE}_{\mathrm{FTE}_{\mathcal{P},t},\boldsymbol{w}_{\mathcal{P}}}(F,\boldsymbol{y}) &= \sum_{P\in\mathcal{P}} w_P \mathrm{SE}_{\mathrm{FTE}_{P,t}}(F,\boldsymbol{y}); \\
&= \sum_{P\in\mathcal{P}} w_P \mathrm{SE}\big(\mathrm{FTE}_{P,t}(F), \mathrm{FTE}_{P,t}(\boldsymbol{y})\big) \\
&= \sum_{P\in\mathcal{P}} w_P \big(\mathbb{E}_F[\mathrm{FTE}_{P,t}(X)] - \mathrm{FTE}_{P,t}(\boldsymbol{y})\big)^2
\end{aligned}$$

where $\mathcal{P}$ is an ensemble of spatial patches, w_P is the weight associated with a patch $P \in \mathcal{P}$ and $\mathrm{FTE}_{P,t}$ the fraction of threshold exceedance over the patch P and for the threshold t (4.18). This scoring rule is proper and focuses on the prediction of the exceedance of a threshold t via the fraction of locations over a patch P exceeding said threshold. The resemblance with the Brier score is clear and the aggregated SE of FTE becomes the aggregated BS when patches of

size $s = 1$ are considered.

In Figure 4.4, the values of the aggregated SE of FTE have been obtained by sampling the forecasts' distribution. Figure 4.4a compares the aggregated CRPS and the aggregated CRPS of spatial mean for different patch size s. For all the scoring rules, we observe an increase in the expected value with the increase of the range of the noise r. As expected, the aggregated CRPS is very sensitive to noise in the mean or the variance and, thus, is prone to the double-penalty effect. The aggregated CRPS of spatial mean is less sensitive to noise on the mean or the variance. Moreover, different patch sizes allow us to select the spatial scale below which we want to avoid a double penalty. Given that the distribution of the noise is fixed in this simulation (i.e., uniform), patch size is related to the level of random fluctuations (i.e., the range r) tolerated by the scoring rule before significant discrimination with respect to the ideal forecast. It is worth noting that the range r of the noise leads to a stronger increase in the values of these CRPS-related scoring rules when the noise is on the mean rather than on the variance.

Figure 4.4b compares the aggregated BS and the aggregated squared error of fraction of threshold exceedances. For simplicity, we fix the threshold $t = 1$. The aggregated BS is, as expected, sensitive to noise in the mean or the variance, and an increase in the range of the noise leads to an increase in the expected value of the score. The aggregated SE of FTE acts as a natural extension of the aggregated BS to patches and provides scoring rules that are less sensitive to noise on the mean or the variance. The sensitivity evolves differently with the increase of the patch size s compared to the aggregated CRPS of spatial mean since the aggregated SE of FTE measures the effect on the average exceedance over a patch. The range r of the noise apparently leads to a comparable increase in the values of the aggregated SE of FTE when the noise is additive or multiplicative.

The use of transformations over patches is similar to neighborhood-based methods in the spatial verification tools framework. Even though avoiding the double-penalty effect is not restricted to tools that do not penalize forecasts below a certain scale, this simulation setup presents a type of test relevant to probability forecasts. The patched-based scoring rules proposed here are not by themselves a sufficient verification tool since they are insensitive to some unrealistic forecast (e.g., if the mean value over the patch is correct but deviations may be as large as possible and lead to unchanged values of the scoring rule). As for any other scoring rule, they should be used with other scoring rules.

4.6 Conclusion

Verification of probabilistic forecasts is an essential but complex step of all forecasting procedures. Scoring rules may appear as the perfect tool to compare forecast performance since, when proper, they can simultaneously assess calibration and sharpness. However, propriety, even if strict, does not ensure that a scoring rule is relevant to the problem at hand. With that in mind, we agree with the recommendation of Scheuerer and Hamill (2015b) that "several different scores be always considered before drawing conclusions". This is even more important in a multivariate setting where forecasts are characterized by more complex objects.

We proposed a framework to construct proper scoring rules in a multivariate setting using aggregation and transformation principles. Aggregation-and-transformation-based scoring rules can improve the conclusions drawn since they enable the verification of specific aspects of the forecast (e.g., anisotropy of the dependence structure). This has been illustrated both using examples from the literature and numerical experiments showcasing different settings. Moreover, we showed that the aggregation and transformation principles can be used to construct scoring rules that are proper, interpretable, and not affected by the double-penalty effect. This could

be a starting point to help bridging the gap between the proper scoring rule community and the spatial verification tools community.

As the interest for machine learning-based weather forecast is increasing (see, e.g., Ben Bouallègue et al. 2024a), multiple approaches have performance comparable to ECMWF deterministic high-resolution forecasts (Keisler, 2022; Pathak et al., 2022; Bi et al., 2023; Lam et al., 2023; Chen et al., 2023). The natural extension to probabilistic forecast is already developing and enabled by publicly available benchmark datasets such as WeatherBench 2 (Rasp et al., 2024). Aggregation-and-transformation-based methods can help ensure that parameter inference does not hedge certain important aspects of the multivariate probabilistic forecasts.

There seems to be a trade-off between discrimination ability and strict propriety. Discrimination ability comes from the ability of scoring rules to differentiate misspecification of certain characteristics. By definition, the expectation of strictly proper scoring rules is minimized when the probabilistic forecast is the true distribution. Nonetheless, it does not guarantee that this global minimum is steep in any misspecification direction. However, interpretable scoring rules can discriminate the misspecification of their target characteristic. Should scoring rules discriminating any misspecification be pursued? Or should interpretable scoring rules discriminating a specific type of misspecification be used instead?

Acknowledgments

The authors acknowledge the support of the French Agence Nationale de la Recherche (ANR) under reference ANR-20-CE40-0025-01 (T-REX project) and the Energy-oriented Centre of Excellence II (EoCoE-II), Grant Agreement 824158, funded within the Horizon2020 framework of the European Union. Part of this work was also supported by the ExtremesLearning grant from 80 PRIME CNRS-INSU and this study has received funding from Agence Nationale de la Recherche - France 2030 as part of the PEPR TRACCS program under grant number ANR-22-EXTR-0005 and the ANR EXSTA.
Sam Allen is thanked for fruitful discussions during the preparation of this manuscript.

4.7 Appendix

4.7.1 Expected univariate scoring rules

Squared Error

For any $F, G \in \mathcal{P}_2(\mathbb{R})$, the expectation of the squared error (4.2) is :

$$\mathbb{E}_G[\mathrm{SE}(F, Y)] = (\mu_F - \mu_G)^2 + {\sigma_G}^2,$$

where μ_F is the mean of the distribution F and μ_G and ${\sigma_G}^2$ are the mean and the variance of the distribution G.

Proof.

$$\begin{aligned}\mathbb{E}_G[\mathrm{SE}(F, Y)] &= \mathbb{E}_G[(\mu_F - Y)^2] \\ &= \mu_F^2 - 2\ \mu_F \mathbb{E}_G[Y] + \mathbb{E}_G[Y^2]\end{aligned}$$

Using the fact that $\mathbb{E}[X^2] = \mathrm{Var}(X) + \mathbb{E}[X]^2$,

$$\begin{aligned}\mathbb{E}_G[\mathrm{SE}(F, Y)] &= \mu_F^2 - 2\ \mu_F \mu_G + \sigma_G^2 + \mu_G^2 \\ &= (\mu_F - \mu_G)^2 + \sigma_G^2\end{aligned}$$

□

Quantile Score

For any $F, G \in \mathcal{P}_1(\mathbb{R})$, the expectation of the quantile score of level α (4.4) is :

$$\begin{aligned}\mathbb{E}_G[\mathrm{QS}_\alpha(F,Y)] &= \int_{-\infty}^{F^{-1}(\alpha)} (F^{-1}(\alpha) - y)G(\mathrm{d}y) - \alpha \int_{\mathbb{R}} (F^{-1}(\alpha) - y)G(\mathrm{d}y); \\ &= \mathbb{E}_G[\mathrm{QS}_\alpha(G,Y)] + \left\{ (G(F^{-1}(\alpha)) - \alpha)(F^{-1}(\alpha) - G^{-1}(\alpha)) - \int_{G^{-1}(\alpha)}^{F^{-1}(\alpha)} (y - G^{-1}(\alpha))G(\mathrm{d}y) \right\}.\end{aligned}$$

Proof. Inspired by the proof of the propriety of the quantile score in Friederichs and Hense (2008).

$$\begin{aligned}\mathbb{E}_G[\mathrm{QS}_\alpha(F,Y)] &= \int_{\mathbb{R}} (\mathbb{1}_{y \le F^{-1}(\alpha)} - \alpha)(F^{-1}(\alpha) - y)G(\mathrm{d}y) \\ &= \int_{-\infty}^{F^{-1}(\alpha)} (1-\alpha)(F^{-1}(\alpha) - y)G(\mathrm{d}y) + \int_{F^{-1}(\alpha)}^{+\infty} (-\alpha)(F^{-1}(\alpha) - y)G(\mathrm{d}y) \\ &= \int_{-\infty}^{F^{-1}(\alpha)} (F^{-1}(\alpha) - y)G(\mathrm{d}y) - \alpha \int_{\mathbb{R}} (F^{-1}(\alpha) - y)G(\mathrm{d}y)\end{aligned}$$

Then, using $F^{-1}(\alpha) - y = (F^{-1}(\alpha) - G^{-1}(\alpha)) + (G^{-1}(\alpha) - y)$,

$$\begin{aligned}\mathbb{E}_G[\mathrm{QS}_\alpha(F,Y)] &= \int_{-\infty}^{F^{-1}(\alpha)} (F^{-1}(\alpha) - G^{-1}(\alpha))G(\mathrm{d}y) - \alpha \int_{\mathbb{R}} (F^{-1}(\alpha) - G^{-1}(\alpha))G(\mathrm{d}y) \\ &\quad + \int_{-\infty}^{F^{-1}(\alpha)} (G^{-1}(\alpha) - y)G(\mathrm{d}y) - \alpha \int_{\mathbb{R}} (G^{-1}(\alpha) - y)G(\mathrm{d}y) \\ &= (G(F^{-1}(\alpha)) - \alpha)(F^{-1}(\alpha) - G^{-1}(\alpha)) \\ &\quad + \int_{-\infty}^{F^{-1}(\alpha)} (G^{-1}(\alpha) - y)G(\mathrm{d}y) - \alpha \int_{\mathbb{R}} (G^{-1}(\alpha) - y)G(\mathrm{d}y) \\ &= (G(F^{-1}(\alpha)) - \alpha)(F^{-1}(\alpha) - G^{-1}(\alpha)) \\ &\quad + \int_{-\infty}^{G^{-1}(\alpha)} (G^{-1}(\alpha) - y)G(\mathrm{d}y) + \int_{G^{-1}(\alpha)}^{F^{-1}(\alpha)} (G^{-1}(\alpha) - y)G(\mathrm{d}y) - \alpha \int_{\mathbb{R}} (G^{-1}(\alpha) - y)G(\mathrm{d}y) \\ &= (G(F^{-1}(\alpha)) - \alpha)(F^{-1}(\alpha) - G^{-1}(\alpha)) + \mathbb{E}_G[\mathrm{QS}_\alpha(G,Y)]) - \int_{G^{-1}(\alpha)}^{F^{-1}(\alpha)} (y - G^{-1}(\alpha))G(\mathrm{d}y)\end{aligned}$$

□

Absolute Error

First of all, for $F \in \mathcal{P}_1(\mathbb{R})$ and $y \in \mathbb{R}$, the absolute error (4.3) is equal to twice the quantile score of level $\alpha = 0.5$:

$$\mathrm{AE}(F,y) = |\mathrm{med}(F) - y| = 2\ \mathrm{QS}_{0.5}(F,y),$$

where $\mathrm{med}(F)$ is the median of the distribution F.

It can be deduced that, for any $F, G \in \mathcal{P}_1(\mathbb{R})$, the expectation of the absolute error is :

$$\begin{aligned}
\mathbb{E}_G[\mathrm{AE}(F,Y)] &= \mathbb{E}_G[|\mathrm{med}(F) - Y|]; \\
&= 2\int_{-\infty}^{\mathrm{med}(F)} (\mathrm{med}(F) - y)G(\mathrm{d}y) - 2\alpha \int_{\mathbb{R}} (\mathrm{med}(F) - y)G(\mathrm{d}y); \\
&= \mathbb{E}_G[\mathrm{AE}(G,Y)] + 2\left\{ (G(\mathrm{med}(F)) - \alpha)(\mathrm{med}(F) - \mathrm{med}(G)) - \int_{\mathrm{med}(G)}^{\mathrm{med}(F)} (y - \mathrm{med}(G))G(\mathrm{d}y) \right\}.
\end{aligned}$$

Brier score

For any $F, G \in \mathcal{P}(\mathbb{R})$, the expectation of the Brier score (4.5) is :

$$\mathbb{E}_G[\mathrm{BS}_t(F,Y)] = (F(t) - G(t))^2 + G(t)(1 - G(t)).$$

Proof.

$$\begin{aligned}
\mathbb{E}_G[\mathrm{BS}_t(F,Y)] &= \mathbb{E}_G[(F(t) - \mathbb{1}_{Y \leq t})^2] \\
&= F(t)^2 - 2F(t)\mathbb{E}_G[\mathbb{1}_{Y \leq t}] + \mathbb{E}_G[{\mathbb{1}_{Y \leq t}}^2] \\
&= F(t)^2 - 2F(t)G(t) + G(t) \\
&= F(t)^2 - 2F(t)G(t) + G(t)^2 - G(t)^2 + G(t) \\
&= (F(t) - G(t))^2 + G(t)(1 - G(t))
\end{aligned}$$

□

Continuous Ranked Probability Score

For any $F, G \in \mathcal{P}_1(\mathbb{R})$, the expectation of the CRPS (4.7) is :

$$\begin{aligned}
\mathbb{E}_G[\mathrm{CRPS}(F,Y)] &= \mathbb{E}_{F,G}|X - Y| - \frac{1}{2}\mathbb{E}_F|X - X'|; \\
&= \int_{\mathbb{R}} (F(z) - G(z))^2 \mathrm{d}z + \int_{\mathbb{R}} G(z)(1 - G(z))\mathrm{d}z,
\end{aligned}$$

where the second term of the last line is the entropy of the CRPS.

Proof.

$$\begin{aligned}
\mathbb{E}_G[\mathrm{CRPS}(F,Y)] &= \mathbb{E}_G\left[\int_{\mathbb{R}} (F(z) - \mathbb{1}_{y \leq z})^2 \mathrm{d}z\right] \\
&= \int_{\mathbb{R}} \mathbb{E}_G\left[(F(z) - \mathbb{1}_{y \leq z})^2\right] \mathrm{d}z \\
&= \int_{\mathbb{R}} \mathbb{E}_G\left[F(z)^2 - 2F(z)\mathbb{1}_{y \leq z} + \mathbb{1}_{y \leq z}^2\right] \mathrm{d}z \\
&= \int_{\mathbb{R}} \left\{F(z)^2 - 2F(z)\mathbb{E}_G\left[\mathbb{1}_{y \leq z}\right] + \mathbb{E}_G\left[\mathbb{1}_{y \leq z}\right]\right\} \mathrm{d}z \\
&= \int_{\mathbb{R}} \left\{F(z)^2 - 2F(z)G(z) + G(z)\right\} \mathrm{d}z \\
&= \int_{\mathbb{R}} \left\{F(z)^2 - 2F(z)G(z) + G(z)^2 - G(z)^2 + G(z)\right\} \mathrm{d}z \\
&= \int_{\mathbb{R}} (F(z) - G(z))^2 \mathrm{d}z + \int_{\mathbb{R}} G(z)(1 - G(z))\mathrm{d}z
\end{aligned}$$

□

Dawid-Sebastiani score

For any $F, G \in \mathcal{P}_2(\mathbb{R})$, the expectation of the Dawid-Sebastiani score (4.9) is :

$$\mathbb{E}_G[\mathrm{DSS}(F,Y)] = \frac{(\mu_F - \mu_G)^2}{{\sigma_F}^2} + \frac{{\sigma_G}^2}{{\sigma_F}^2} + 2\log\sigma_F.$$

Proof.

$$\begin{aligned}\mathbb{E}_G[\mathrm{DSS}(F,Y)] &= \mathbb{E}_G\left[\frac{(Y-\mu_F)^2}{{\sigma_F}^2} + 2\log\sigma_F\right] \\ &= \frac{\mathbb{E}_G\left[(Y-\mu_F)^2\right]}{{\sigma_F}^2} + 2\log\sigma_F\end{aligned}$$

Noticing that $\mathbb{E}_G\left[(Y-\mu_F)^2\right] = \mathbb{E}_G\left[\mathrm{SE}(F,Y)\right]$,

$$\mathbb{E}_G[\mathrm{DSS}(F,Y)] = \frac{(\mu_F - \mu_G)^2 + {\sigma_G}^2}{{\sigma_F}^2} + 2\log\sigma_F.$$

□

Error-spread score

For any $F, G \in \mathcal{P}_4(\mathbb{R})$, the expectation of the error-spread score (4.10) is :

$$\begin{aligned}\mathbb{E}_G[\mathrm{ESS}(F,Y)] = &\left[({\sigma_G}^2 - {\sigma_F}^2) + (\mu_G - \mu_F)^2 - \sigma_F\gamma_F(\mu_G - \mu_F)\right]^2 \\ &+ {\sigma_G}^2\left[2(\mu_G - \mu_F) + (\sigma_G\gamma_G - \sigma_F\gamma_F)\right]^2 \\ &+ {\sigma_G}^4(\beta_G - {\gamma_G}^2 - 1),\end{aligned}$$

where μ_F, σ_F^2, γ_F are the mean, the variance and the skewness of the probabilistic forecast F. Similarly, μ_G, σ_G^2, γ_G and β_G are the first four centered moments of the distribution G. The proof is available in Appendix B of Christensen et al. (2014).

Logarithmic score

For any $F, G \in \mathcal{P}(\mathbb{R})$ such that F and G have probability density functions in the class $\mathcal{L}_1(\mathbb{R})$, the expectation of the logarithmic score (4.11) is :

$$\mathbb{E}_G[\mathrm{LogS}(F,Y)] = D_{\mathrm{KL}}(G||F) + \mathrm{H}(F),$$

where $D_{\mathrm{KL}}(G||F)$ is the Kullback-Leibler divergence from F to G and $\mathrm{H}(F)$ is the Shannon entropy of F. The proof is straightforward given that the Kullback-Leibler divergence and Shannon entropy are defined as

$$\begin{aligned}D_{\mathrm{KL}}(G||F) &= \int_{\mathbb{R}} g(y)\log\left(\frac{g(y)}{f(y)}\right)\mathrm{d}y; \\ \mathrm{H}(F) &= \int_{\mathbb{R}} f(y)\log(f(y))\mathrm{d}y.\end{aligned}$$

Hyvärinen score

For F, G such that their densities f exist, are twice continuously differentiable and satisfy $f'(x)/f(x) \to 0$ as $|x| \to \infty$ and $g'(x)/g(x) \to 0$ as $|x| \to \infty$, the expectation of the Hyvärinen score is :

$$\begin{aligned}\mathbb{E}_G[\mathrm{HS}(F,Y)] &= \int_{\mathbb{R}}\left(\frac{f'(y)^2}{f(y)^2} - 2\frac{f'(y)g'(y)}{f(y)g(y)}\right)g(y)\mathrm{d}y \\ &= \int_{\mathbb{R}}\left(\frac{f'(y)}{f(y)} - \frac{g'(y)}{g(y)}\right)^2 g(y)\mathrm{d}y - \int_{\mathbb{R}}\frac{g'(y)^2}{g(y)^2}g(y)\mathrm{d}y\end{aligned}$$

where the last formula shows the entropy of the Hyvärinen score (second term on the right-hand side).

Proof.

$$\begin{aligned}\mathbb{E}_G[\mathrm{HS}(F,Y)] &= \mathbb{E}\left[2\frac{f''(Y)}{f(Y)} - \frac{f'(Y)^2}{f(y)^2}\right] \\ &= 2\int_{\mathbb{R}} \frac{f''(y)}{f(y)} g(y)\mathrm{d}y - \int_{\mathbb{R}} \frac{f'(y)^2}{f(y)^2} g(y)\mathrm{d}y\end{aligned}$$

Integrating by part the integral of the first term on the right-hand side leads to :

$$\begin{aligned}\int_{\mathbb{R}} \frac{f''(y)}{f(y)} g(y)\mathrm{d}y &= \left[\frac{f'(y)}{f(y)} g(y)\right]_{-\infty}^{+\infty} - \int_{\mathbb{R}} f'(y) \frac{g'(y)f(y) - g(y)f'(y)}{f(y)^2}\mathrm{d}y \\ &= -\int_{\mathbb{R}} \frac{f'(y)g'(y)}{f(y)g(y)} g(y)\mathrm{d}y + \int_{\mathbb{R}} \frac{f'(y)^2}{f(y)^2} g(y)\mathrm{d}y\end{aligned}$$

The boundary term is null since $f'(x)/f(x) \to 0$ as $|x| \to \infty$ and g is a probability density function.
Thus,

$$\begin{aligned}\mathbb{E}_G[\mathrm{HS}(F,Y)] &= -2\int_{\mathbb{R}} \frac{f'(y)g'(y)}{f(y)g(y)} g(y)\mathrm{d}y + 2\int_{\mathbb{R}} \frac{f'(y)^2}{f(y)^2} g(y)\mathrm{d}y - \int_{\mathbb{R}} \frac{f'(y)^2}{f(y)^2} g(y)\mathrm{d}y \\ &= -2\int_{\mathbb{R}} \frac{f'(y)g'(y)}{f(y)g(y)} g(y)\mathrm{d}y + \int_{\mathbb{R}} \frac{f'(y)^2}{f(y)^2} g(y)\mathrm{d}y \\ &= \int_{\mathbb{R}} \left(\frac{f'(y)^2}{f(y)^2} - 2\frac{f'(y)g'(y)}{f(y)g(y)}\right) g(y)\mathrm{d}y\end{aligned}$$

□

Quadratic score

For any $F, G \in \mathcal{L}_2(\mathbb{R})$, the expectation of the quadratic score is :

$$\mathbb{E}_G[\mathrm{QuadS}(F,Y)] = \|f\|_2^2 - 2\langle f, g\rangle,$$

where $\langle f, g\rangle = \int_{\mathbb{R}} f(y)g(y)\mathrm{d}y$.

Pseudospherical score

For any $F, G \in \mathcal{L}_\alpha(\mathbb{R})$, the expectation of the quadratic score is :

$$\mathbb{E}_G[\mathrm{PseudoS}(F,Y)] = -\frac{\langle f^{\alpha-1}, g\rangle}{\|f\|_\alpha^{\alpha-1}},$$

where $\langle f^{\alpha-1}, g\rangle = \int_{\mathbb{R}} f(y)^{\alpha-1} g(y)\mathrm{d}y$.

4.7.2 Expected multivariate scoring rules

Squared error

For any $F, G \in \mathcal{P}_2(\mathbb{R}^d)$, the expectation of the squared error (4.12) is :

$$\mathbb{E}_G[\mathrm{SE}(F, \boldsymbol{Y})] = \|\boldsymbol{\mu}_F - \boldsymbol{\mu}_G\|_2^2 + \mathrm{tr}(\Sigma_G),$$

where $\boldsymbol{\mu}_F$ is the mean vector of the distribution F and $\boldsymbol{\mu}_G$ and ${\Sigma_G}^2$ are the mean vector and the covariance matrix of the distribution G.

Proof. Let T_i denote the projection on the i-th margin.

$$\begin{aligned}
\mathbb{E}_G[\mathrm{SE}(F, \boldsymbol{Y})] &= \mathbb{E}_G[\|\boldsymbol{\mu}_F - \boldsymbol{Y}\|_2^2] \\
&= \mathbb{E}_G\left[\sum_{i=1}^{d} (\boldsymbol{\mu}_{T_i(F)} - T_i(\boldsymbol{Y}))^2\right] \\
&= \sum_{i=1}^{d} \mathbb{E}_{T_i(G)}\left[\mathrm{SE}(T_i(F), Y)\right] \\
&= \sum_{i=1}^{d} \left((\mu_{T_i(F)} - \mu_{T_i(G)})^2 + \sigma^2_{T_i(G)}\right) \\
&= \|\boldsymbol{\mu}_F - \boldsymbol{\mu}_G\|_2^2 + \mathrm{tr}(\Sigma_G)
\end{aligned}$$

□

Dawid-Sebastiani score

For any $F, G \in \mathcal{P}_2(\mathbb{R}^d)$, the expectation of the Dawid-Sebastiani score is :

$$\mathbb{E}_G[\mathrm{DSS}(F, \boldsymbol{Y})] = \log(\det \Sigma_F) + (\boldsymbol{\mu}_F - \boldsymbol{\mu}_G)^T \Sigma_F^{-1} (\boldsymbol{\mu}_F - \boldsymbol{\mu}_G) + \mathrm{tr}(\Sigma_G \Sigma_F^{-1}).$$

The proof is available in the original article (Dawid and Sebastiani, 1999).

Energy score

In a general setting, the expected energy score does not simplify. For any $F, G \in \mathcal{P}_\beta(\mathbb{R}^d)$, the expected energy score (4.13) is :

$$\mathbb{E}_G[\mathrm{ES}_\beta(F, \boldsymbol{Y})] = \mathbb{E}_{F,G}\|\boldsymbol{X} - \boldsymbol{Y}\|_2^\beta - \frac{1}{2}\mathbb{E}_F\|\boldsymbol{X} - \boldsymbol{X}'\|_2^\beta.$$

Variogram score

For any $F, G \in \mathcal{P}(\mathbb{R}^d)$ such that the $2p$-th moments of all their univariate margins are finite, the expected variogram score of order p (4.14) is :

$$\mathbb{E}_G[\mathrm{VS}_p(F, \boldsymbol{Y})] = \sum_{i,j=1}^{d} w_{ij} \left(\mathbb{E}_F\left[|X_i - X_j|^p\right]^2 - 2\mathbb{E}_F\left[|X_i - X_j|^p\right] \mathbb{E}_G[|Y_i - Y_j|^p] + \mathbb{E}_G[|Y_i - Y_j|^{2p}]\right).$$

Proof.

$$\begin{aligned}
\mathbb{E}_G[\mathrm{VS}_p(F, \boldsymbol{Y})] &= \mathbb{E}_G\left[\sum_{i,j=1}^{d} w_{ij}\left(\mathbb{E}_F\left[|X_i - X_j|^p\right] - |Y_i - Y_j|^p\right)^2\right] \\
&= \mathbb{E}_G\left[\sum_{i,j=1}^{d} w_{ij}\left(\mathbb{E}_F\left[|X_i - X_j|^p\right]^2 - 2\mathbb{E}_F\left[|X_i - X_j|^p\right]|Y_i - Y_j|^p + |Y_i - Y_j|^{2p}\right)\right] \\
&= \sum_{i,j=1}^{d} w_{ij}\left(\mathbb{E}_F\left[|X_i - X_j|^p\right]^2 - 2\mathbb{E}_F\left[|X_i - X_j|^p\right]\mathbb{E}_G[|Y_i - Y_j|^p] + \mathbb{E}_G[|Y_i - Y_j|^{2p}]\right).
\end{aligned}$$

□

Logarithmic score

For any $F, G \in \mathcal{P}(\mathbb{R}^d)$ such that F and G have probability density functions that belong to $\mathcal{L}_1(\mathbb{R}^d)$, the expectation of the logarithmic score is analogous to its univariate version :

$$\mathbb{E}_G[\mathrm{LogS}(F, \boldsymbol{Y})] = D_{\mathrm{KL}}(G||F) + \mathrm{H}(F),$$

where $D_{\mathrm{KL}}(G||F)$ is the Kullback-Leibler divergence from F to G and $\mathrm{H}(F)$ is the Shannon entropy of F.

$$\begin{aligned}
D_{\mathrm{KL}}(G||F) &= \int_{\mathbb{R}^d} g(\boldsymbol{y}) \log\left(\frac{g(\boldsymbol{y})}{f(\boldsymbol{y})}\right) \mathrm{d}\boldsymbol{y} \\
\mathrm{H}(F) &= \int_{\mathbb{R}^d} f(\boldsymbol{y}) \log(f(\boldsymbol{y}))\mathrm{d}\boldsymbol{y}.
\end{aligned}$$

Hyvärinen score

For $F, G \in \mathcal{P}(\mathbb{R}^d)$ such that their probability density functions f and g such that they are twice continuously differentiable and satisfying $\nabla f(x) \to 0$ and $\nabla g(x) \to 0$ as $\|x\| \to \infty$, the expectation of the Hyvärinen score is :

$$\mathbb{E}[\mathrm{HS}(F, \boldsymbol{Y})] = \int_{\mathbb{R}^d} g(y)\langle \nabla \log(f(y)) - 2\nabla \log(g(y)), \nabla \log(f(y))\rangle g(y)\mathrm{d}y$$

where ∇ is the gradient operator and $\langle \cdot, \cdot \rangle$ is the scalar product. The proof is similar to the proof for the univariate case using integration by parts and Stoke's theorem (Parry et al., 2012).

Quadratic score

For any $F, G \in \mathcal{L}_2(\mathbb{R}^d)$, the expectation of the quadratic score is analogous to its univariate version :

$$\mathbb{E}_G[\mathrm{QuadS}(F, \boldsymbol{Y})] = \|f\|_2^2 - 2\langle f, g\rangle,$$

where $\langle f, g\rangle = \int_{\mathbb{R}^d} f(\boldsymbol{y})g(\boldsymbol{y})\mathrm{d}\boldsymbol{y}$.

Pseudospherical score

For any $F, G \in \mathcal{L}_\alpha(\mathbb{R}^d)$, the expectation of the quadratic score is analogous to its univariate version :

$$\mathbb{E}_G[\mathrm{PseudoS}(F, \boldsymbol{Y})] = -\frac{\langle f^{\alpha-1}, g\rangle}{\|f\|_\alpha^{\alpha-1}},$$

where $\langle f^{\alpha-1}, g\rangle = \int_{\mathbb{R}^d} f(\boldsymbol{y})^{\alpha-1}g(\boldsymbol{y})\mathrm{d}\boldsymbol{y}$.

4.7.3 Proofs

Proposition 4.1

Proof of Proposition 4.1. Let $\mathcal{F} \subset \mathcal{P}(\mathbb{R}^d)$ be a class of Borel probability measure on $\mathbb{R}^d$ and let $F \in \mathcal{F}$ be a forecast and $y \in \mathbb{R}^d$ an observation. Let $T : \mathbb{R}^d \to \mathbb{R}^k$ be a transformation and let S be a scoring rule on $\mathbb{R}^k$ that is proper relative to $T(\mathcal{F}) = \{\mathcal{L}(T(\boldsymbol{X})), X \sim F \in \mathcal{F}\}$.

$$\begin{aligned}\mathbb{E}_G\left[\mathrm{S}_T(F, \boldsymbol{Y})\right] &= \mathbb{E}_G\left[\mathrm{S}(T(F)), T(\boldsymbol{Y}))\right] \\ &= \mathbb{E}_{T(G)}\left[\mathrm{S}(T(F), \boldsymbol{Y})\right]\end{aligned}$$

Given that $T(F), T(G) \in T(\mathcal{F})$ and S is proper relative to $T(\mathcal{F})$,

$$\begin{aligned}&\mathbb{E}_{T(G)}\left[\mathrm{S}(T(G), \boldsymbol{Y})\right] \leq \mathbb{E}_{T(G)}\left[\mathrm{S}(T(F), \boldsymbol{Y})\right] \\ \Leftrightarrow\ &\mathbb{E}_G\left[\mathrm{S}_T(G, \boldsymbol{Y})\right] \leq \mathbb{E}_G\left[\mathrm{S}_T(F, \boldsymbol{Y})\right] \qquad (4.23)\end{aligned}$$

□

Proof of the strict propriety case in Proposition 4.1. The notations are the same as the proof above except the following. Let $T : \mathbb{R}^d \to \mathbb{R}^k$ be an injective transformation and let S be a scoring rule on $\mathbb{R}^k$ that is strictly proper relative to $T(\mathcal{F}) = \{\mathcal{L}(T(\boldsymbol{X})), X \sim F \in \mathcal{F}\}$.

The equality in Equation (4.23) leads to :

$$\begin{aligned}&\mathbb{E}_G\left[\mathrm{S}_T(G, \boldsymbol{Y})\right] = \mathbb{E}_G\left[\mathrm{S}_T(F, \boldsymbol{Y})\right] \\ \Leftrightarrow\ &\mathbb{E}_G\left[\mathrm{S}(T(G), T(\boldsymbol{Y}))\right] = \mathbb{E}_G\left[\mathrm{S}(T(F), T(\boldsymbol{Y}))\right] \\ \Leftrightarrow\ &\mathbb{E}_{T(G)}\left[\mathrm{S}(T(G), \boldsymbol{Y})\right] = \mathbb{E}_{T(G)}\left[\mathrm{S}(T(F), \boldsymbol{Y})\right]\end{aligned}$$

The fact that S is strictly proper relative to $T(\mathcal{F})$ leads to $T(F) = T(G)$, and finally since T is injective, we have $F = G$. □

Proposition 4.3

Proof of Proposition 4.3. The proof relies on the reproducing kernel Hilbert space (RKHS) representation of the kernel scoring rule S_ρ. For a background on kernel scoring rule, maximum mean discrepancies and RKHS, we refer to Smola et al. (2007) or Steinwart and Christmann (2008, Section 4).

Let $\mathcal{H}_\rho$ denote the RKHS associated with ρ. We recall that $\mathcal{H}_\rho$ contains all the functions $\rho(\boldsymbol{x}, \cdot)$ and that the inner product on $\mathcal{H}_\rho$ satisfies the property

$$\langle \rho(\boldsymbol{x}_1, \cdot), \rho(\boldsymbol{x}_2, \cdot) \rangle_{\mathcal{H}_\rho} = \rho(\boldsymbol{x}_1, \boldsymbol{x}_2).$$

The *kernel mean embedding* is a linear application $\Psi_\rho : \mathcal{P}_\rho \to \mathcal{H}_\rho$ mapping an admissible distribution $F \in \mathcal{P}_\rho$ into a function $\Psi_\rho(F)$ in the RKHS and such that the image of the point measure $\delta_{\boldsymbol{x}}$ is $\rho(\boldsymbol{x}, \cdot)$. Equation (4.16) giving the kernel scoring rule for an ensemble prediction $F = \frac{1}{M}\sum_{m=1}^{M} \delta_{\boldsymbol{x}_m}$ can be written as

$$\begin{aligned}\mathrm{S}_\rho(F, \boldsymbol{y}) &= \frac{1}{2}\langle \Psi_\rho(F) - \Psi_\rho(\delta_{\boldsymbol{y}}), \Psi_\rho(F) - \Psi_\rho(\delta_{\boldsymbol{y}}) \rangle_{\mathcal{H}_\rho} \\ &= \frac{1}{2}\|\Psi_\rho(F - \delta_{\boldsymbol{y}})\|^2_{\mathcal{H}_\rho}.\end{aligned}$$

The properties of the kernel mean embedding ensure that this relation still holds for all $F \in \mathcal{P}_\rho$. As a consequence, if $(T_l)_{l\geq 1}$ is an Hilbertian basis of $\mathcal{H}_\rho$, we have

$$\begin{aligned}\mathrm{S}_\rho(F, y) &= \frac{1}{2}\|\Psi_\rho(F - \delta_{\boldsymbol{y}})\|^2_{\mathcal{H}_\rho} \\ &= \frac{1}{2}\sum_{l\geq 1}\langle \Psi_\rho(F - \delta_{\boldsymbol{y}}), T_l\rangle^2_{\mathcal{H}_\rho}.\end{aligned}$$

Finally, the properties of the kernel mean embedding ensure that, for all $T \in \mathcal{H}_\rho$,

$$\langle \Psi_\rho(F - \delta_{\boldsymbol{y}}), T\rangle_{\mathcal{H}_\rho} = \int_{\mathbb{R}^d} T(\boldsymbol{x})(F - \delta_{\boldsymbol{y}})(\mathrm{d}\boldsymbol{x}) = \mathbb{E}_F[T(\boldsymbol{X})] - T(\boldsymbol{y})$$

whence the result follows. □

Proof of examples illustrating Proposition 4.3

Next, we illustrate the Proposition 4.3 and provide some computations in two cases: the Gaussian kernel scoring rule and the continuous rank probability score (CRPS).

Gaussian Kernel Scoring Rule. This is the scoring rule related to the Gaussian kernel

$$\rho(x_1, x_2) = \exp(-(x_1 - x_2)^2/2), \quad x_1, x_2 \in \mathbb{R}.$$

Using a series expansion of the exponential function, we have

$$\rho(x_1, x_2) = \mathrm{e}^{-x_1^2/2}\mathrm{e}^{-x_2^2/2}\sum_{l\geq 0}\frac{(x_1x_2)^l}{l!} = \sum_{l\geq 0} T_l(x_1)T_l(x_2)$$

with T_l the transformation defined, for $l \geq 0$, by

$$T_l(x) = \frac{1}{\sqrt{l!}}\mathrm{e}^{-x^2/2}x^l.$$

As a consequence, the Gaussian kernel scoring rule writes, for all $F \in \mathcal{P}(\mathbb{R})$ and $y \in \mathbb{R}$,

$$\begin{aligned}\mathrm{S}_\rho(F, y) &= \frac{1}{2}\int_{\mathbb{R}\times\mathbb{R}} \rho(x_1, x_2)(F - \delta_y)(\mathrm{d}x_1)(F - \delta_y)(\mathrm{d}x_2) \\ &= \frac{1}{2}\int_{\mathbb{R}\times\mathbb{R}} \Big(\sum_{l\geq 0} T_l(x_1)T_l(x_2)\Big)(F - \delta_y)(\mathrm{d}x_1)(F - \delta_y)(\mathrm{d}x_2) \\ &= \frac{1}{2}\sum_{l\geq 0}\Big(\int_{\mathbb{R}} T_l(x)(F - \delta_y)(\mathrm{d}x)\Big)^2 \\ &= \frac{1}{2}\sum_{l\geq 0}\Big(\mathbb{E}_F[T_l(X)] - T_l(y)\Big)^2.\end{aligned}$$

Continuous Ranked Probability Score. The CRPS is the scoring rule with kernel $\rho(x_1, x_2) = |x_1| + |x_2| - |x_1 - x_2|$. This kernel is the covariance of the Brownian motion on $\mathbb{R}$ and its RKHS is known to be the Sobolev space $\mathrm{H}^1 = \mathrm{H}^1(\mathbb{R})$, see Berlinet and Thomas-Agnan (2004). We recall the definition of the Sobolev space

$$\mathrm{H}^1 = \Big\{f \in \mathcal{C}(\mathbb{R}, \mathbb{R})\colon f(0) = 0, \dot{f} \in L^2(\mathbb{R})\Big\},$$

where $\dot{f}$ denotes the derivative of f assumed to be defined almost everywhere and square-integrable. The inner product on H^1 is defined by

$$\langle f_1, f_2 \rangle_{H^1} = \int_{\mathbb{R}} \dot{f}_1(x)\dot{f}_2(x)\mathrm{d}x$$

and one can easily check the fundamental relation

$$\langle \rho(x_1, \cdot), \rho(x_2, \cdot) \rangle_{H^1} = \int_{\mathbb{R}} \dot{\rho}(x_1, x)\dot{\rho}(x_2, x)\mathrm{d}x = \rho(x_1, x_2).$$

Here the derivative $\dot{\rho}(x_1, x) = \mathbb{1}_{[0,x_1]}(x)$ is taken with respect to the second variable x. Then, we consider the Haar system defined as the collection of functions

$$H_l^0(x) = H^0(x - l) \quad \text{and} \quad H_{l,m}^1(x) = 2^{m/2}H^1(2^m x - l), \quad l \in \mathbb{Z},\, m \geq 0,$$

with $H^0(x) = \mathbb{1}_{[0,1)}(x)$ and $H^1(x) = \mathbb{1}_{[0,1/2)}(x) - \mathbb{1}_{[1/2,1)}(x)$. Since the Haar system is an orthonormal basis of the space $L^2(\mathbb{R})$ and the map $f \in H^1 \mapsto \dot{f} \in L^2$ is an isomorphism between Hilbert spaces, we obtain an orthonormal basis of $\mathrm{H}^1(\mathbb{R})$ by considering the primitives vanishing at 0 of the Haar basis functions. Setting $T^0(x) = x\mathbb{1}_{[0,1)}(x) + \mathbb{1}_{[1,+\infty)}(x)$ and $T^1(x) = \big(1/2 - |x - 1/2|\big)\mathbb{1}_{[0,1]}(x)$ the primitive functions of H^0 and H^1 respectively, we obtain the system

$$T_l^0(x) = T^0(x - l), \quad T_{l,m}^1(x) = 2^{-m/2}T^1(2^m x - l), \quad l \in \mathbb{Z},\, m \geq 0.$$

The series representation of the CRPS is then deduced from Proposition 4.3 and its proof since the collection $\{T_{l,m} \colon l \in \mathbb{Z}, m \geq 0\}$, is an orthonormal basis of the RKHS associated with the kernel ρ of the CRPS.

4.7.4 General form of Corollary 4.1

Corollary 4.2. *Let $\mathcal{T} = \{T_i\}_{1 \leq i \leq m}$ be a set of transformations from $\mathbb{R}^d$ to $\mathbb{R}^k$. Let $\mathcal{S} = \{\mathrm{S}_i\}_{1 \leq i \leq m}$ be a set of proper scoring rules such that S_i is proper relative to $T_i(\mathcal{F})$, for all $1 \leq i \leq m$. Let $\boldsymbol{w} = \{w_i\}_{1 \leq i \leq m}$ be nonnegative weights. Then the scoring rule*

$$\mathrm{S}_{\mathcal{S}\mathcal{T},\boldsymbol{w}}(F, \boldsymbol{y}) = \sum_{i=1}^{m} w_i \mathrm{S}_{iT_i}(F, \boldsymbol{y}) = \sum_{i=1}^{m} w_i \mathrm{S}_i(T_i(F), T_i(\boldsymbol{y}))$$

is proper relative to $\mathcal{F}$.

4.7.5 Scoring rules of the simulation study

The following formulas are deduced for a probabilistic forecast F taking the form of the Gaussian random field model of Equation (4.20). The formulas of the aggregated univariate scoring rules can be obtained from the formulas in Gneiting and Raftery (2007) and Jordan et al. (2019) and, thus, are not presented here. We focus on the expression of the variogram score and the CRPS of spatial mean.

Variogram Score

$$\mathrm{VS}_p(F, \boldsymbol{y}) = \sum_{s,s' \in \mathcal{D}} w_{ss'} \left(\mathbb{E}_F[|X_s - X_{s'}|^p] - |y_s - y_{s'}|^p\right)^2$$

For $X \sim \mathcal{N}(\mu, \sigma^2)$, the absolute moment is (Winkelbauer, 2014) :

$$\mathbb{E}[|X|^\nu] = \sigma^\nu 2^{\nu/2} \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi}} {}_1F_1\left(-\nu/2, 1/2; -\frac{\mu^2}{2\sigma^2}\right), \tag{4.24}$$

where ${}_1F_1$ is the confluent hypergeometric function of the first kind. For $X \sim F$,

$$\begin{aligned} X_s - X_{s'} &\sim \mathcal{N}(\mu_s - \mu_{s'}, {\sigma_s}^2 + {\sigma_{s'}}^2 - 2\text{cov}(F_s, F_{s'}) \\ &\sim \mathcal{N}(0, 2\sigma^2(1 - e^{-\left(\frac{\|s-s'\|}{\lambda}\right)^\beta})). \end{aligned}$$

This leads to

$$\begin{aligned} \mathbb{E}_G[|X_s - X_{s'}|^p] &= \left(2\sigma^2(1 - e^{-\left(\frac{\|s-s'\|}{\lambda}\right)^\beta})\right)^{p/2} 2^{p/2}\frac{\Gamma\left(\frac{p+1}{2}\right)}{\sqrt{\pi}} {}_1F_1\left(-p/2, 1/2; -\frac{(\mu_s - \mu_{s'})^2}{4\sigma^2(1 - e^{-\left(\frac{\|s-s'\|}{\lambda}\right)^\beta})}\right) \\ &= 2^p\sigma^p\left(1 - e^{-\left(\frac{\|s-s'\|}{\lambda}\right)^\beta}\right)^{p/2} \frac{\Gamma\left(\frac{p+1}{2}\right)}{\sqrt{\pi}} {}_1F_1\left(-p/2, 1/2; 0\right) \\ &= 2^p\sigma^p\left(1 - e^{-\left(\frac{\|s-s'\|}{\lambda}\right)^\beta}\right)^{p/2} \frac{\Gamma\left(\frac{p+1}{2}\right)}{\sqrt{\pi}} \end{aligned}$$

Finally,

$$\begin{aligned} \text{VS}_p(F, \boldsymbol{y}) &= \sum_{s,s' \in \mathcal{D}} w_{ij} \left(\mathbb{E}_G[|X_s - X_{s'}|^p] - |y_s - y_{s'}|^p\right)^2 \\ &= \sum_{s,s' \in \mathcal{D}} w_{ij} \left(\left(2\sigma^2(1 - e^{-\left(\frac{\|s-s'\|}{\lambda}\right)^\beta})\right)^{p/2} 2^{p/2}\frac{\Gamma\left(\frac{p+1}{2}\right)}{\sqrt{\pi}} - |y_s - y_{s'}|^p\right)^2 \end{aligned}$$

p-Variation Score

$$\begin{aligned} \text{pVS}(F, \boldsymbol{y}) &= \sum_{\boldsymbol{s} \in \mathcal{D}^*} w_{\boldsymbol{s}} \text{SE}_{T_{p-var,\boldsymbol{s}}}(F, \boldsymbol{y}); \\ &= \sum_{\boldsymbol{s} \in \mathcal{D}^*} w_{\boldsymbol{s}} (\mathbb{E}_F[T_{p-var,\boldsymbol{s}}(\boldsymbol{X})] - T_{p-var,\boldsymbol{s}}(\boldsymbol{y}))^2, \end{aligned}$$

Denote $Z = \boldsymbol{X}_{\boldsymbol{s}+(1,1)} - \boldsymbol{X}_{\boldsymbol{s}+(1,0)} - \boldsymbol{X}_{\boldsymbol{s}+(0,1)} + \boldsymbol{X}_{\boldsymbol{s}}$. For $X \sim F$, we have $Z \sim \mathcal{N}(\mu_Z, \sigma_Z^2)$ with

$$\mu_Z = \mu_{\boldsymbol{s}+(1,1)} - \mu_{\boldsymbol{s}+(1,0)} - \mu_{\boldsymbol{s}+(0,1)} + \mu_{\boldsymbol{s}} = 0$$

and

$$\begin{aligned} \sigma_Z^2 &= \sigma^2_{\boldsymbol{s}+(1,1)} + \sigma^2_{\boldsymbol{s}+(1,0)} + \sigma^2_{\boldsymbol{s}+(0,1)} + \sigma^2_{\boldsymbol{s}} \\ &\quad - 2\text{cov}(F(\boldsymbol{s}+(1,1)), F(\boldsymbol{s}+(1,0))) - 2\text{cov}(F(\boldsymbol{s}+(1,1)), F(\boldsymbol{s}+(0,1)) + 2\text{cov}(F(\boldsymbol{s}+(1,1)), F(\boldsymbol{s})) \\ &\quad + 2\text{cov}(F(\boldsymbol{s}+(1,0)), F(\boldsymbol{s}+(0,1))) - 2\text{cov}(F(\boldsymbol{s}+(1,0)), F(\boldsymbol{s})) \\ &\quad - 2\text{cov}(F(\boldsymbol{s}+(0,1)), F(\boldsymbol{s})) \\ &= 4\sigma^2(1 + e^{-(\sqrt{2}/\lambda)^\beta} - 2e^{-(1/\lambda)^\beta}) \end{aligned}$$

Using (4.24), this leads to

$$\begin{aligned} \mathbb{E}_F[T_{p-var,\boldsymbol{s}}(\boldsymbol{X})] &= \left(4\sigma^2(1 + e^{-(\sqrt{2}/\lambda)^\beta} - 2e^{-(1/\lambda)^\beta})\right)^{p/2} 2^{p/2}\frac{\Gamma\left(\frac{p+1}{2}\right)}{\sqrt{\pi}} {}_1F_1\left(-p/2, 1/2; 0\right) \\ &= \left(4\sigma^2(1 + e^{-(\sqrt{2}/\lambda)^\beta} - 2e^{-(1/\lambda)^\beta})\right)^{p/2} 2^{p/2}\frac{\Gamma\left(\frac{p+1}{2}\right)}{\sqrt{\pi}} \end{aligned}$$

Finally,

$$\begin{aligned}\text{pVS}(F, \boldsymbol{y}) &= \sum_{\boldsymbol{s}\in\mathcal{D}^*} w_{\boldsymbol{s}} \text{SE}_{T_{p-var,\boldsymbol{s}}}(F, \boldsymbol{y}) \\ &= \sum_{\boldsymbol{s}\in\mathcal{D}^*} w_{\boldsymbol{s}} \Bigg(\left(4\sigma^2(1 + e^{-(\sqrt{2}/\lambda)^\beta} - 2e^{-(1/\lambda)^\beta})\right)^{p/2} 2^{p/2} \frac{\Gamma\left(\frac{p+1}{2}\right)}{\sqrt{\pi}} \\ &\qquad - |y_{\boldsymbol{s}+(1,1)} - y_{\boldsymbol{s}+(1,0)} - y_{\boldsymbol{s}+(0,1)} + y_{\boldsymbol{s}}|^p \Bigg)^2\end{aligned}$$

CRPS of spatial mean

The CRPS of spatial mean is defined as

$$\begin{aligned}\text{CRPS}_{\text{mean}_{\mathcal{P}},\boldsymbol{w}_{\boldsymbol{\mathcal{P}}}}(F, \boldsymbol{y}) &= \sum_{P\in\mathcal{P}} w_P \text{CRPS}_{\text{mean}_P}(F, \boldsymbol{y}) \\ &= \sum_{P\in\mathcal{P}} w_P \text{CRPS}(\text{mean}_P(F), \text{mean}_P(\boldsymbol{y})),\end{aligned}$$

where $\mathcal{P}$ is an ensemble of spatial patches and w_P is the weight associated with a patch $P \in \mathcal{P}$. The mean of Gaussian marginals follows a Gaussian distribution :

$$\text{mean}_P(F) \sim \mathcal{N}(\sum_{s\in P} \mu_s, \frac{\sigma^2}{|P|^2} \sum_{s,s'\in P} e^{-(\frac{\|s-s'\|}{\lambda})^\beta}) = \mathcal{N}(\mu_P, \sigma_P^2),$$

where $|P|$ is the cardinal of the patch P (i.e., the number of grid points belonging to P).

Finally,

$$\text{CRPS}_{\text{mean}_{\mathcal{P}},\boldsymbol{w}_{\boldsymbol{\mathcal{P}}}}(F, \boldsymbol{y}) = \sum_{P\in\mathcal{P}} w_P \text{CRPS}(\mathcal{N}(\mu_P, \sigma_P^2), \text{mean}_P(\boldsymbol{y})).$$

Chapter 5

Perspectives

Chapter 2 provides optimal minimax convergence rates and upper bounds in distributional regression for a theoretical risk based on the CRPS. A first natural extension of this work would be to obtain similar results for other scoring rules. However, as discussed in Appendix C, the current form of the proofs might be ill-suited to certain scoring rules (e.g., logarithmic score; Roulston and Smith 2002) and would require to be adapted by using other algorithms than k-NN and uniform kernels. If the proofs of Chapter 2 can be obtained using the kernel form of the CRPS, this could represent an entry point to kernel scores (Steinwart and Ziegel, 2021) and, in particular, multivariate distributional regression. Another possible perspective is investigating whether state-of-the-art statistical postprocessing methods achieve optimal minimax convergence rates. The most direct extension seems to be neighborhood-based methods such as quantile regression forests (QRF; Meinshausen 2006), used as a benchmark method in Chapter 3. However, this would again require modifying the proofs since QRF uses weights depending on neighborhoods based on both covariables and the target variable. It would also be interesting to investigate the rate of convergence of neural network methods such as distributional neural networks (Rasp and Lerch, 2018). Last, it could be of interest to focus on universal convergence, as done in Appendix A, for a large class of methods (e.g., distributional random forests or distributional neural networks) and for a large class of scoring rules. Universal convergence studies the convergence on the broadest class of distributions without guarantee on the convergence rate.

One of the known empirical limitations of analog ensembles is that they generally require more training data than other postprocessing techniques to achieve the same level of predictive performance (see, e.g., Taillardat et al. 2019). As mentioned in Chapter 3, large and consistent training datasets are not always available due to NWP updates and the computational cost of reforecasts. In this regard, the optimal use of the information contained in a finite training dataset is crucial in an operational setting. A possible perspective could be to investigate non-asymptotical upper bounds on the theoretical risk of state-of-the-art statistical postprocessing methods to obtain guarantees on their use of information. This could unveil a better understanding of the empirically observed difference in training dataset size requirement between analog ensembles and QRF. For example, it may originate from their different dependence on the intrinsic characteristic of the true distribution (e.g., associated with its dispersion or its regularity) rather than directly on the size of the training dataset. The operational limited consistency of training datasets could also be investigated theoretically by relaxing the independent and identically distributed hypothesis.

In addition to the rise of machine learning (ML-)based statistical postprocessing methods mentioned in Chapter 3, the increase of ML-based methods affects weather forecasting in general (Ben Bouallègue et al., 2024a). Regarding deterministic forecasts, multiple approaches have performance comparable to ECMWF deterministic high-resolution forecasts (Keisler, 2022; Pathak

et al., 2022; Bi et al., 2023; Lam et al., 2023; Chen et al., 2023). Most of these models learn to minimize root mean square error, which tends to lead to blurry forecasts for large lead times. Similar effects can appear for probabilistic forecasts due to the double-penalty effect (Ebert, 2008). As recalled in Chapter 4, an effort to develop spatial verification tools has been carried out to avoid this undesired effect and specifically assess different important characteristics of forecasts (Gilleland et al., 2009; Dorninger et al., 2018). The development of similar tools for probabilistic forecasts within the framework of proper scoring rules will be crucial for both the development and the verification of ML-based weather forecasts. Pacchiardi et al. (2024) have exhibited promising results by proposing a generative method for weather forecasting based on the minimization of the patched energy score. We may hope that our contributions from Chapter 4 to the theory of spatial scoring rules will be useful for future developments in ML-based weather forecasting.

Bibliography

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL https://www.tensorflow.org/. Software available from tensorflow.org.

Paolo Agnolucci, Chrysanthi Rapti, Peter Alexander, Vincenzo De Lipsis, Robert A. Holland, Felix Eigenbrod, and Paul Ekins. Impacts of rising temperatures and farm management practices on global yields of 18 crops. *Nature Food*, 1(9):562–571, 2020. ISSN 2662-1355. https://doi.org/10.1038/s43016-020-00148-x.

Zeina Al Masry, Romain Pic, Clément Dombry, and Chrisine Devalland. A new methodology to predict the oncotype scores based on clinico-pathological data with similar tumor profiles. *Breast Cancer Research and Treatment*, 203(3):587–598, 2023. ISSN 1573-7217. https://doi.org/10.1007/s10549-023-07141-5.

Kathy S Albain, William E Barlow, Steven Shak, Gabriel N Hortobagyi, Robert B Livingston, I-Tien Yeh, Peter Ravdin, Roberto Bugarini, Frederick L Baehner, Nancy E Davidson, George W Sledge, Eric P Winer, Clifford Hudis, James N Ingle, Edith A Perez, Kathleen I Pritchard, Lois Shepherd, Julie R Gralow, Carl Yoshizawa, D Craig Allred, C Kent Osborne, and Daniel F Hayes. Prognostic and predictive value of the 21-gene recurrence score assay in postmenopausal women with node-positive, oestrogen-receptor-positive breast cancer on chemotherapy: a retrospective analysis of a randomised trial. *The Lancet Oncology*, 11(1): 55–65, 2010. ISSN 1470-2045. https://doi.org/10.1016/s1470-2045(09)70314-6.

Joan Albanell, Christer Svedman, Joseph Gligorov, Simon D.H. Holt, Gianfilippo Bertelli, Jens-Uwe Blohmer, Roman Rouzier, Ana Lluch, and Wolfgang Eiermann. Pooled analysis of prospective european studies assessing the impact of using the 21-gene recurrence score assay on clinical decision making in women with oestrogen receptor–positive, human epidermal growth factor receptor 2–negative early-stage breast cancer. *European Journal of Cancer*, 66: 104–113, 2016. ISSN 0959-8049. https://doi.org/10.1016/j.ejca.2016.06.027.

C. Alexander, M. Coulon, Y. Han, and X. Meng. Evaluating the discrimination ability of proper multi-variate scoring rules. *Annals of Operations Research*, 334(1):857–883, 2022. ISSN 1572-9338. https://doi.org/10.1007/s10479-022-04611-9.

Sam Allen, Jonas Bhend, Olivia Martius, and Johanna Ziegel. Weighted verification tools to evaluate univariate and multivariate probabilistic forecasts for high-impact weather events. *Weather and Forecasting*, 38(3):499–516, 2023a. ISSN 1520-0434. https://doi.org/10.1175/waf-d-22-0161.1.

Sam Allen, David Ginsbourger, and Johanna Ziegel. Evaluating forecasts for high-impact events using transformed kernel scores. *SIAM/ASA Journal on Uncertainty Quantification*, 11(3): 906–940, 2023b. ISSN 2166-2525. https://doi.org/10.1137/22m1532184.

Sam Allen, Johanna Ziegel, and David Ginsbourger. Assessing the calibration of multivariate probabilistic forecasts. *Quarterly Journal of the Royal Meteorological Society*, 150(760):1315–1335, 2024. ISSN 1477-870X. https://doi.org/10.1002/qj.4647.

Jeffrey L. Anderson. A method for producing and evaluating probabilistic forecasts from ensemble model integrations. *Journal of Climate*, 9(7):1518–1530, 1996. ISSN 1520-0442. https://doi.org/10.1175/1520-0442(1996)009<1518:amfpae>2.0.co;2.

Fabrice Andre, Nofisat Ismaila, N. Lynn Henry, Mark R. Somerfield, Robert C. Bast, William Barlow, Deborah E. Collyar, M. Elizabeth Hammond, Nicole M. Kuderer, Minetta C. Liu, Catherine Van Poznak, Antonio C. Wolff, and Vered Stearns. Use of biomarkers to guide decisions on adjuvant systemic therapy for women with early-stage invasive breast cancer: Asco clinical practice guideline update—integration of results from tailorx. *Journal of Clinical Oncology*, 37(22):1956–1964, 2019. ISSN 1527-7755. https://doi.org/10.1200/jco.19.00945.

Susan Athey, Julie Tibshirani, and Stefan Wager. Generalized random forests. *The Annals of Statistics*, 47(2):1148 – 1178, 2019. ISSN 0090-5364. https://doi.org/10.1214/18-AOS1709.

Aline Baltres, Zeina Al Masry, Ryad Zemouri, Severine Valmary-Degano, Laurent Arnould, Noureddine Zerhouni, and Christine Devalland. Prediction of oncotype dx recurrence score using deep multi-layer perceptrons in estrogen receptor-positive, her2-negative breast cancer. *Breast Cancer*, 27(5):1007–1016, 2020. ISSN 1880-4233. https://doi.org/10.1007/s12282-020-01100-4.

Sándor Baran and Sebastian Lerch. Combining predictive distributions for the statistical post-processing of ensemble forecasts. *International Journal of Forecasting*, 34(3):477–496, 2018. ISSN 0169-2070. https://doi.org/10.1016/j.ijforecast.2018.01.005.

Sándor Baran and Dóra Nemoda. Censored and shifted gamma distribution based emos model for probabilistic quantitative precipitation forecasting. *Environmetrics*, 27(5):280–292, 2016. ISSN 1099-095X. https://doi.org/10.1002/env.2391.

Andreas Basse-O'Connor, Vytautė Pilipauskaitė, and Mark Podolskij. Power variations for fractional type infinitely divisible random fields. *Electronic Journal of Probability*, 26:1 – 35, 2021. https://doi.org/10.1214/21-EJP617.

Peter W. Battaglia, Jessica B. Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, Caglar Gulcehre, Francis Song, Andrew Ballard, Justin Gilmer, George Dahl, Ashish Vaswani, Kelsey Allen, Charles Nash, Victoria Langston, Chris Dyer, Nicolas Heess, Daan Wierstra, Pushmeet Kohli, Matt Botvinick, Oriol Vinyals, Yujia Li, and Razvan Pascanu. Relational inductive biases, deep learning, and graph networks. 2018. https://doi.org/10.48550/ARXIV.1806.01261.

Peter Bauer, Alan Thorpe, and Gilbert Brunet. The quiet revolution of numerical weather prediction. *Nature*, 525(7567):47–55, 2015. https://doi.org/10.1038/nature14956.

Erhan Bayraktar and Gaoyue Guo. Strong equivalence between metrics of wasserstein type. *Electronic Communications in Probability*, 26:1 – 13, 2021. ISSN 1083-589X. https://doi.org/10.1214/21-ECP383.

Zied Ben Bouallègue, Tobias Heppelmann, Susanne E. Theis, and Pierre Pinson. Generation of scenarios from calibrated ensemble forecasts with a dual-ensemble copula-coupling approach. *Monthly Weather Review*, 144(12):4737–4750, 2016. ISSN 1520-0493. https://doi.org/10.1175/mwr-d-15-0403.1.

Zied Ben Bouallègue, Mariana C. A. Clare, Linus Magnusson, Estibaliz Gascón, Michael Maier-Gerber, Martin Janoušek, Mark Rodwell, Florian Pinault, Jesper S. Dramsch, Simon T. K. Lang, Baudouin Raoult, Florence Rabier, Matthieu Chevallier, Irina Sandu, Peter Dueben, Matthew Chantry, and Florian Pappenberger. The rise of data-driven weather forecasting: A first statistical assessment of machine learning–based weather forecasts in an operational-like context. *Bulletin of the American Meteorological Society*, 105(6):E864–E883, 2024a. ISSN 1520-0477. https://doi.org/10.1175/bams-d-23-0162.1.

Zied Ben Bouallègue, Jonathan A. Weyn, Mariana C. A. Clare, Jesper Dramsch, Peter Dueben, and Matthew Chantry. Improving medium-range ensemble weather forecasts with hierarchical ensemble transformers. *Artificial Intelligence for the Earth Systems*, 3(1):e230027, 2024b. ISSN 2769-7525. https://doi.org/10.1175/AIES-D-23-0027.1.

Albert Benassi, Serge Cohen, and Jacques Istas. On roughness indices for fractional fields. *Bernoulli*, 10(2):357 – 373, 2004. https://doi.org/10.3150/bj/1082380223.

Patrick Bénichou. Cartography of statistical pluviometric fields with an automatic allowance for regional topography. In *Global Precipitations and Climate Change*, pages 187–199. Springer Berlin Heidelberg, 1994. https://doi.org/10.1007/978-3-642-79268-7_11.

Alain Berlinet and Christine Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Kluwer Academic Publishers, Boston, MA, 2004. ISBN 1-4020-7679-7. https://doi.org/10.1007/978-1-4419-9096-9. With a preface by Persi Diaconis.

Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Accurate medium-range global weather forecasting with 3d neural networks. *Nature*, 619(7970):533–538, 2023. ISSN 1476-4687. https://doi.org/10.1038/s41586-023-06185-3.

Gérard Biau and Luc Devroye. *Lectures on the Nearest Neighbor Method*. Springer Series in the Data Sciences. Springer, 2015.

Vilhelm Bjerknes. The problem of weather prediction, considered from the viewpoints of mechanics and physics. *Meteorologische Zeitschrift*, 18(6):663–667, 2009. ISSN 0941-2948. https://doi.org/10.1127/0941-2948/2009/416.

Mathias Blicher Bjerregård, Jan Kloppenborg Møller, and Henrik Madsen. An introduction to multivariate probabilistic forecast evaluation. *Energy and AI*, 4:100058, 2021. ISSN 2666-5468. https://doi.org/10.1016/j.egyai.2021.100058.

David Bolin and Jonas Wallin. Local scale invariance and robustness of proper scoring rules. *Statistical Science*, 38(1):140 – 159, 2023. https://doi.org/10.1214/22-STS864.

Nikos I. Bosse, Sam Abbott, Anne Cori, Edwin van Leeuwen, Johannes Bracher, and Sebastian Funk. Scoring epidemiological forecasts on transformed scales. *PLOS Computational Biology*, 19(8):1–23, 2023. ISSN 1553-7358. https://doi.org/10.1371/journal.pcbi.1011393.

François Bouttier, Laure Raynaud, Olivier Nuissier, and Benjamin Ménétrier. Sensitivity of the AROME ensemble to initial and surface perturbations during HyMeX. *Quarterly Journal of the Royal Meteorological Society*, 142(S1):390–403, 2015. https://doi.org/10.1002/qj.2622.

Jonas Brehmer. Elicitability and its application in risk management. Master's thesis, University of Mannheim, 2017.

Jonas R. Brehmer and Kirstin Strokorb. Why scoring functions cannot assess tail properties. *Electronic Journal of Statistics*, 13(2):4015 – 4034, 2019. ISSN 1935-7524. https://doi.org/10.1214/19-EJS1622.

Leo Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996. ISSN 1573-0565. https://doi.org/10.1007/bf00058655.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. ISSN 1573-0565. https://doi.org/10.1023/a:1010933404324.

Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. *Classification And Regression Trees*. Taylor & Francis, 1984. ISBN 9780412048418. https://doi.org/10.1201/9781315139470.

John Bjørnar Bremnes. Ensemble postprocessing using quantile function regression based on neural networks and bernstein polynomials. *Monthly Weather Review*, 148(1):403–414, 2020. ISSN 1520-0493. https://doi.org/10.1175/MWR-D-19-0227.1.

Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950. ISSN 1520-0493. https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

Jochen Bröcker. Reliability, sufficiency, and the decomposition of proper scores. *Quarterly Journal of the Royal Meteorological Society*, 135(643):1512–1519, 2009. ISSN 1477-870X. https://doi.org/10.1002/qj.456.

Jochen Bröcker and Zied Ben Bouallègue. Stratified rank histograms for ensemble forecast verification under serial dependence. *Quarterly Journal of the Royal Meteorological Society*, 146(729):1976–1990, 2020. ISSN 1477-870X. https://doi.org/10.1002/qj.3778.

Jochen Bröcker and Leonard A. Smith. Scoring probabilistic forecasts: The importance of being proper. *Weather and Forecasting*, 22(2):382–388, 2007. ISSN 0882-8156. https://doi.org/10.1175/waf966.1.

Roberto Buizza. *Ensemble Forecasting and the Need for Calibration*, pages 15–48. Elsevier, 2018. ISBN 9780128123720. https://doi.org/10.1016/b978-0-12-812372-0.00002-9.

Sebastian Buschow. Measuring displacement errors with complex wavelets. *Weather and Forecasting*, 37(6):953–970, 2022. ISSN 1520-0434. https://doi.org/10.1175/waf-d-21-0180.1.

Sebastian Buschow and Petra Friederichs. Using wavelets to verify the scale structure of precipitation forecasts. *Advances in Statistical Climatology, Meteorology and Oceanography*, 6 (1):13–30, 2020. ISSN 2364-3587. https://doi.org/10.5194/ascmo-6-13-2020.

Sebastian Buschow and Petra Friederichs. Sad: Verifying the scale, anisotropy and direction of precipitation forecasts. *Quarterly Journal of the Royal Meteorological Society*, 147(735): 1150–1169, 2021. ISSN 1477-870X. https://doi.org/10.1002/qj.3964.

Sebastian Buschow, Jakiw Pidstrigach, and Petra Friederichs. Assessment of wavelet-based spatial verification by means of a stochastic precipitation model (wv_verif v0.1.0). *Geoscientific Model Development*, 12(8):3401–3418, 2019. ISSN 1991-9603. https://doi.org/10.5194/gmd-12-3401-2019.

Barbara Casati, Manfred Dorninger, Caio A. S. Coelho, Elizabeth E. Ebert, Chiara Marsigli, Marion P. Mittermaier, and Eric Gilleland. The 2020 international verification methods workshop online: Major outcomes and way forward. *Bulletin of the American Meteorological Society*, 103(3):E899–E910, 2022. ISSN 1520-0477. https://doi.org/10.1175/bams-d-21-0126.1.

Olivier Caumont, Marc Mandement, François Bouttier, Judith Eeckman, Cindy Lebeaupin Brossier, Alexane Lovat, Olivier Nuissier, and Olivier Laurantin. The heavy precipitation event of 14–15 october 2018 in the aude catchment: a meteorological study based on operational numerical weather prediction systems and standard and personal observations. *Natural Hazards and Earth System Sciences*, 21(3):1135–1157, 2021. ISSN 1684-9981. https://doi.org/10.5194/nhess-21-1135-2021.

Jean-Louis Champeaux, Pascale Dupuy, Olivier Laurantin, Isabelle Soulan, Pierre Tabary, and Jean-Michel Soubeyroux. Les mesures de précipitations et l'estimation des lames d'eau à Météo-France : état de l'art et perspectives. *La Houille Blanche*, 95(5):28–34, 2009. ISSN 1958-5551. https://doi.org/10.1051/lhb/2009052.

William E. Chapman, Luca Delle Monache, Stefano Alessandrini, Aneesh C. Subramanian, F. Martin Ralph, Shang-Ping Xie, Sebastian Lerch, and Negin Hayatbini. Probabilistic predictions from deterministic atmospheric river forecasts with deep learning. *Monthly Weather Review*, 150(1):215–234, 2022. ISSN 1520-0493. https://doi.org/10.1175/mwr-d-21-0106.1.

Gaoxiang Chen, Qun Li, Fuqian Shi, Islem Rekik, and Zhifang Pan. RFDCR: Automated brain lesion segmentation using cascaded random forests with dense conditional random fields. *NeuroImage*, 211:116620, 2020. ISSN 1053-8119. https://doi.org/10.1016/j.neuroimage.2020.116620.

Kang Chen, Tao Han, Junchao Gong, Lei Bai, Fenghua Ling, Jing-Jia Luo, Xi Chen, Leiming Ma, Tianning Zhang, Rui Su, Yuanzheng Ci, Bin Li, Xiaokang Yang, and Wanli Ouyang. Fengwu: Pushing the skillful global medium-range weather forecast beyond 10 days lead. 2023. https://doi.org/10.48550/ARXIV.2304.02948.

François Chollet et al. Keras, 2015. URL https://keras.io.

H. M. Christensen, I. M. Moroz, and T. N. Palmer. Evaluation of ensemble forecast uncertainty using a new proper score: Application to medium-range and seasonal forecasts. *Quarterly Journal of the Royal Meteorological Society*, 141(687):538–549, 2014. ISSN 1477-870X. https://doi.org/10.1002/qj.2375.

Daniela Cisneros, Jordan Richards, Ashok Dahal, Luigi Lombardo, and Raphaël Huser. Deep graphical regression for jointly moderate and extreme australian wildfires. *Spatial Statistics*, 59:100811, 2024. ISSN 2211-6753. https://doi.org/10.1016/j.spasta.2024.100811.

Martyn Clark, Subhrendu Gangopadhyay, Lauren Hay, Balaji Rajagopalan, and Robert Wilby. The schaake shuffle: A method for reconstructing space–time variability in forecasted precipitation and temperature fields. *Journal of Hydrometeorology*, 5(1):243–262, 2004. https://doi.org/10.1175/1525-7541(2004)005<0243:tssamf>2.0.co;2.

T. Cover and P. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27, 1967. ISSN 1557-9654. https://doi.org/10.1109/tit.1967.1053964.

Y. Dai and S. Hemri. Spatially coherent postprocessing of cloud cover ensemble forecasts. *Monthly Weather Review*, 149(12):3923–3937, 2021. https://doi.org/10.1175/mwr-d-21-0046.1.

A. P. Dawid. Present position and potential developments: Some personal views: Statistical theory: The prequential approach. *Journal of the Royal Statistical Society. Series A (General)*, 147(2):278–292, 1984. ISSN 0035-9238. https://doi.org/10.2307/2981683.

A. Philip Dawid and Paola Sebastiani. Coherent dispersion criteria for optimal experimental design. *The Annals of Statistics*, 27(1):65 – 81, 1999. ISSN 0090-5364. https://doi.org/10.1214/aos/1018031101.

A. Philip Dawid, Monica Musio, and Laura Ventura. Minimum scoring rule inference. *Scandinavian Journal of Statistics*, 43(1):123–138, 2015. ISSN 1467-9469. https://doi.org/10.1111/sjos.12168.

Alexander Philip Dawid and Monica Musio. Theory and applications of proper scoring rules. *METRON*, 72(2):169–183, 2014. ISSN 2281-695X. https://doi.org/10.1007/s40300-014-0039-y.

Luca Delle Monache, F. Anthony Eckel, Daran L. Rife, Badrinath Nagarajan, and Keith Searight. Probabilistic weather prediction with an analog ensemble. *Monthly Weather Review*, 141(10):3498–3516, 2013. ISSN 1520-0493. https://doi.org/10.1175/mwr-d-12-00281.1.

Francis X. Diebold and Roberto S. Mariano. Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3):253–263, 1995. ISSN 1537-2707. https://doi.org/10.1080/07350015.1995.10524599.

Jean Diebolt, Armelle Guillou, and Imen Rached. Approximation of the distribution of excesses through a generalized probability-weighted moments method. *Journal of Statistical Planning and Inference*, 137(3):841–857, 2007. https://doi.org/10.1016/j.jspi.2006.06.012.

Clément Dombry, Thibault Modeste, and Romain Pic. Stone's theorem for distributional regression in wasserstein distance. *Journal of Nonparametric Statistics*, pages 1–23, 2024. ISSN 1029-0311. https://doi.org/10.1080/10485252.2024.2393172.

Manfred Dorninger, Eric Gilleland, Barbara Casati, Marion P. Mittermaier, Elizabeth E. Ebert, Barbara G. Brown, and Laurence J. Wilson. The setup of the mesovict project. *Bulletin of the American Meteorological Society*, 99(9):1887–1906, 2018. ISSN 1520-0477. https://doi.org/10.1175/bams-d-17-0164.1.

Elizabeth E. Ebert. Fuzzy verification of high-resolution gridded forecasts: a review and proposed framework. *Meteorological Applications*, 15(1):51–64, 2008. https://doi.org/10.1002/met.25.

Werner Ehm and Tilmann Gneiting. Local proper scoring rules of order two. *The Annals of Statistics*, 40(1):609 – 637, 2012. ISSN 0090-5364. https://doi.org/10.1214/12-AOS973.

Kira Feldmann, David S. Richardson, and Tilmann Gneiting. Grid- versus station-based postprocessing of ensemble temperature forecasts. *Geophysical Research Letters*, 46(13):7744–7751, 2019. ISSN 1944-8007. https://doi.org/10.1029/2019gl083189.

Carlos Fernandez-Lozano, Pablo Hervella, Virginia Mato-Abad, Manuel Rodríguez-Yáñez, Sonia Suárez-Garaboa, Iria López-Dequidt, Ana Estany-Gestal, Tomás Sobrino, Francisco Campos, José Castillo, Santiago Rodríguez-Yáñez, and Ramón Iglesias-Rey. Random forest-based prediction of stroke outcome. *Scientific Reports*, 11(1):10071, 2021. ISSN 2045-2322. https://doi.org/10.1038/s41598-021-89434-7.

C. A. T. Ferro. Fair scores for ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*, 140(683):1917–1923, 2013. ISSN 0035-9009. https://doi.org/10.1002/qj.2270.

Christopher A. T. Ferro, David S. Richardson, and Andreas P. Weigel. On the effect of ensemble size on the discrete and continuous ranked probability scores. *Meteorological Applications*, 15 (1):19–24, 2008. ISSN 1469-8080. https://doi.org/10.1002/met.45.

Melina B Flanagan, David J Dabbs, Adam M Brufsky, Sushil Beriwal, and Rohit Bhargava. Histopathologic variables predict oncotype dx™ recurrence score. *Modern Pathology*, 21(10): 1255–1261, 2008. ISSN 0893-3952. https://doi.org/10.1038/modpathol.2008.54.

Nicolas Fournier and Arnaud Guillin. On the rate of convergence in wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3–4):707–738, 2015. ISSN 1432-2064. https://doi.org/10.1007/s00440-014-0583-7.

Petra Friederichs and Andreas Hense. A probabilistic forecast approach for daily precipitation totals. *Weather and Forecasting*, 23(4):659–673, 2008. ISSN 0882-8156. https://doi.org/10.1175/2007waf2007051.1.

Petra Friederichs and Thordis L. Thorarinsdottir. Forecast verification for extreme value distributions with an application to probabilistic peak wind prediction. *Environmetrics*, 23(7): 579–594, 2012. ISSN 1180-4009. https://doi.org/10.1002/env.2176.

Petra Friederichs, Sabrina Wahl, and Sebastian Buschow. Postprocessing for extreme events. In *Statistical Postprocessing of Ensemble Forecasts*, pages 127–154. Elsevier, 2018. https://doi.org/10.1016/b978-0-12-812372-0.00005-4.

Urs Germann, Marco Boscacci, Lorenzo Clementi, Marco Gabella, Alessandro Hering, Maurizio Sartori, Ioannis V. Sideris, and Bertrand Calpini. Weather radar in complex orography. *Remote Sensing*, 14(3):503, 2022. ISSN 2072-4292. https://doi.org/10.3390/rs14030503.

Eric Gilleland. Spatial forecast verification: Baddeley's delta metric applied to the icp test cases. *Weather and Forecasting*, 26(3):409–415, 2011. ISSN 1520-0434. https://doi.org/10.1175/waf-d-10-05061.1.

Eric Gilleland, David Ahijevych, Barbara G. Brown, Barbara Casati, and Elizabeth E. Ebert. Intercomparison of spatial forecast verification methods. *Weather and Forecasting*, 24(5): 1416–1430, 2009. ISSN 0882-8156. https://doi.org/10.1175/2009waf2222269.1.

Armando E. Giuliano, James L. Connolly, Stephen B. Edge, Elizabeth A. Mittendorf, Hope S. Rugo, Lawrence J. Solin, Donald L. Weaver, David J. Winchester, and Gabriel N. Hortobagyi. Breast cancer—major changes in the american joint committee on cancer eighth edition cancer staging manual. *CA: A Cancer Journal for Clinicians*, 67(4):290–303, 2017. ISSN 1542-4863. https://doi.org/10.3322/caac.21393.

Tilmann Gneiting. Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106(494):746–762, 2011. ISSN 1537-274X. https://doi.org/10.1198/jasa.2011.r10138.

Tilmann Gneiting. On the cover-hart inequality: What's a sample of size one worth? *Stat*, 1 (1):12–17, 2012. ISSN 2049-1573. https://doi.org/10.1002/sta4.3.

Tilmann Gneiting and Matthias Katzfuss. Probabilistic forecasting. *Annual Review of Statistics and Its Application*, 1(1):125–151, 2014. ISSN 2326-831X. https://doi.org/10.1146/annurev-statistics-062713-085831.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. ISSN 1537-274X. https://doi.org/10.1198/016214506000001437.

Tilmann Gneiting and Roopesh Ranjan. Comparing density forecasts using threshold- and quantile-weighted scoring rules. *Journal of Business and Economic Statistics*, 29(3):411–422, 2011. ISSN 1537-2707. https://doi.org/10.1198/jbes.2010.08110.

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. ISSN 0027-0644. https://doi.org/10.1175/mwr2904.1.

Tilmann Gneiting, Fadoua Balabdaoui, and Adrian E. Raftery. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 69(2):243–268, 2007. ISSN 1467-9868. https://doi.org/10.1111/j.1467-9868.2007.00587.x.

Tilmann Gneiting, Larissa I. Stanberry, Eric P. Grimit, Leonhard Held, and Nicholas A. Johnson. Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds. *TEST*, 17(2):211–235, 2008. ISSN 1863-8260. https://doi.org/10.1007/s11749-008-0114-x.

Tilmann Gneiting, Sebastian Lerch, and Benedikt Schulz. Probabilistic solar forecasting: Benchmarks, post-processing, verification. *Solar Energy*, 252:72–80, 2023. ISSN 0038-092X. https://doi.org/10.1016/j.solener.2022.12.054.

I. J. Good. Rational decisions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 14(1):107–114, 1952. ISSN 2517-6161. https://doi.org/10.1111/j.2517-6161.1952.tb00104.x.

J. Arthur Greenwood, J. Maciunas Landwehr, N. C. Matalas, and J. R. Wallis. Probability weighted moments: Definition and relation to parameters of several distributions expressable in inverse form. *Water Resources Research*, 15(5):1049–1054, 1979. ISSN 1944-7973. https://doi.org/10.1029/wr015i005p01049.

Peter Grönquist, Chengyuan Yao, Tal Ben-Nun, Nikoli Dryden, Peter Dueben, Shigang Li, and Torsten Hoefler. Deep learning for post-processing ensemble weather forecasts. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 379 (2194):20200092, 2021. https://doi.org/10.1098/rsta.2020.0092.

Lászlò Györfi, Michael Kohler, Adam Krzyzak, and Harro Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer Series in Statistics. Springer, 2002.

Thomas M. Hamill. *Practical Aspects of Statistical Postprocessing*, pages 187–217. Elsevier, 2018. ISBN 9780128123720. https://doi.org/10.1016/b978-0-12-812372-0.00007-8.

Thomas M. Hamill and Stephen J. Colucci. Verification of eta–rsm short-range ensemble forecasts. *Monthly Weather Review*, 125(6):1312–1327, 1997. ISSN 1520-0493. https://doi.org/10.1175/1520-0493(1997)125<1312:VOERSR>2.0.CO;2.

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. ISSN 0027-0644. https://doi.org/10.1175/mwr3237.1.

Fan Han and Istvan Szunyogh. A technique for the verification of precipitation forecasts and its application to a problem of predictability. *Monthly Weather Review*, 146(5):1303–1318, 2018. ISSN 1520-0493. https://doi.org/10.1175/mwr-d-17-0040.1.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York, 2009. ISBN 9780387848587. https://doi.org/10.1007/978-0-387-84858-7.

Claudio Heinrich-Mertsching, Thordis L. Thorarinsdottir, Peter Guttorp, and Max Schneider. Validation of point process predictions with proper scoring rules. *Scandinavian Journal of Statistics*, 2024. ISSN 1467-9469. https://doi.org/10.1111/sjos.12736.

S. Hemri, M. Scheuerer, F. Pappenberger, K. Bogner, and T. Haiden. Trends in the predictive performance of raw ensemble weather forecasts. *Geophysical Research Letters*, 41(24):9197–9205, 2014. https://doi.org/10.1002/2014gl062472.

Alexander Henzi, Johanna F. Ziegel, and Tilmann Gneiting. Isotonic distributional regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 83(5):963–993, 2021. https://doi.org/10.1111/rssb.12450.

Hans Hersbach. Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, 15(5):559–570, 2000. ISSN 1520-0434. https://doi.org/10.1175/1520-0434(2000)015<0559:dotcrp>2.0.co;2.

Hajo Holzmann and Matthias Eulert. The role of the information set for forecasting—with applications to risk management. *The Annals of Applied Statistics*, 8(1):595 – 621, 2014. ISSN 1932-6157. https://doi.org/10.1214/13-AOAS709.

Hajo Holzmann and Bernhard Klar. Focusing on regions of interest in forecast evaluation. *The Annals of Applied Statistics*, 11(4):2404 – 2431, 2017. ISSN 1932-6157. https://doi.org/10.1214/17-AOAS1088.

Nina Horat and Sebastian Lerch. Deep learning for postprocessing global probabilistic forecasts on subseasonal time scales. *Monthly Weather Review*, 152(3):667–687, 2024. ISSN 1520-0493. https://doi.org/10.1175/mwr-d-23-0150.1.

Yanjun Hou, Gary Tozbikian, Debra L. Zynger, and Zaibo Li. Using the modified magee equation to identify patients unlikely to benefit from the 21-gene recurrence score assay (oncotype dx assay). *American Journal of Clinical Pathology*, 147(6):541–548, 2017. ISSN 1943-7722. https://doi.org/10.1093/ajcp/aqx008.

Weiming Hu, Mohammadvaghef Ghazvinian, William E. Chapman, Agniv Sengupta, Fred Martin Ralph, and Luca Delle Monache. Deep learning forecast uncertainty for precipitation over the western united states. *Monthly Weather Review*, 151(6):1367–1385, 2023. ISSN 1520-0493. https://doi.org/10.1175/mwr-d-22-0268.1.

Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(24):695–709, 2005. URL http://jmlr.org/papers/v6/hyvarinen05a.html.

Ian T. Jolliffe and Cristina Primo. Evaluating rank histograms using decompositions of the chi-square test statistic. *Monthly Weather Review*, 136(6):2133–2139, 2008. ISSN 0027-0644. https://doi.org/10.1175/2007mwr2219.1.

Alexander Jordan, Fabian Krüger, and Sebastian Lerch. Evaluating probabilistic forecasts with scoringrules. *Journal of Statistical Software*, 90(12):1–37, 2019. ISSN 1548-7660. https://doi.org/10.18637/jss.v090.i12.

Thomas H. Jordan, Yun-Tai Chen, Paolo Gasparini, Raul Madariaga, Ian Main, Warner Marzocchi, Gerassimos Papadopoulos, Gennady Sobolev, Koshun Yamaoka, and Jochen Zschau. Operational earthquake forecasting. state of knowledge and guidelines for utilization. *Annals of Geophysics*, 54(4), 2011. ISSN 2037-416X. https://doi.org/10.4401/ag-5350.

Victor Richmond Jose. A characterization for the spherical scoring rule. *Theory and Decision*, 66(3):263–281, 2007. ISSN 1573-7187. https://doi.org/10.1007/s11238-007-9067-x.

Kevin Kalinsky, William E. Barlow, Julie R. Gralow, Funda Meric-Bernstam, Kathy S. Albain, Daniel F. Hayes, Nancy U. Lin, Edith A. Perez, Lori J. Goldstein, Stephen K.L. Chia, Sukhbinder Dhesy-Thind, Priya Rastogi, Emilio Alba, Suzette Delaloge, Miguel Martin, Catherine M. Kelly, Manuel Ruiz-Borrego, Miguel Gil-Gil, Claudia H. Arce-Salinas, Etienne G.C. Brain, Eun-Sook Lee, Jean-Yves Pierga, Begoña Bermejo, Manuel Ramos-Vazquez, Kyung-Hae Jung, Jean-Marc Ferrero, Anne F. Schott, Steven Shak, Priyanka Sharma, Danika L. Lew, Jieling Miao, Debasish Tripathy, Lajos Pusztai, and Gabriel N. Hortobagyi. 21-gene assay to inform chemotherapy benefit in node-positive breast cancer. *New England Journal of Medicine*, 385(25):2336–2347, 2021. ISSN 1533-4406. https://doi.org/10.1056/nejmoa2108873.

Ryan Keisler. Forecasting global weather with graph neural networks. 2022. https://doi.org/10.48550/ARXIV.2202.07575.

Isaac Kim, Hee Jun Choi, Jai Min Ryu, Se Kyung Lee, Jong Han Yu, Seok Won Kim, Seok Jin Nam, and Jeong Eon Lee. A predictive model for high/low risk group according to oncotype dx recurrence score using machine learning. *European Journal of Surgical Oncology*, 45(2): 134–140, 2019. ISSN 0748-7983. https://doi.org/10.1016/j.ejso.2018.09.011.

Molly E Klein, David J Dabbs, Yongli Shuai, Adam M Brufsky, Rachel Jankowitz, Shannon L Puhalla, and Rohit Bhargava. Prediction of the oncotype dx recurrence score: use of pathology-generated equations derived by linear regression analysis. *Modern Pathology*, 26 (5):658–664, 2013. ISSN 0893-3952. https://doi.org/10.1038/modpathol.2013.36.

S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951. ISSN 0003-4851. https://doi.org/10.1214/aoms/1177729694.

Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, Alexander Merose, Stephan Hoyer, George Holland, Oriol Vinyals, Jacklynn Stott, Alexander Pritzel, Shakir Mohamed, and Peter Battaglia. Learning skillful medium-range global weather forecasting. *Science*, 382(6677):1416–1421, 2023. ISSN 1095-9203. https://doi.org/10.1126/science.adi2336.

Sebastian Lerch and Kai L. Polsterer. Convolutional autoencoders for spatially-informed ensemble post-processing. In *ICLR 2022 - AI for Earth and Space Science Workshop*, 2022.

Sebastian Lerch and Thordis L. Thorarinsdottir. Comparison of non-homogeneous regression models for probabilistic wind speed forecasting. *Tellus A: Dynamic Meteorology and Oceanography*, 65(1):21206, 2013. ISSN 1600-0870. https://doi.org/10.3402/tellusa.v65i0.21206.

Sebastian Lerch, Thordis L. Thorarinsdottir, Francesco Ravazzolo, and Tilmann Gneiting. Forecaster's dilemma: Extreme events and forecast evaluation. *Statistical Science*, 32(1):106 – 127, 2017. ISSN 0883-4237. https://doi.org/10.1214/16-STS588.

Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Rui Li, Brian J. Reich, and Howard D. Bondell. Deep distribution regression. *Computational Statistics & Data Analysis*, 159:107203, 2021. ISSN 0167-9473. https://doi.org/10.1016/j.csda.2021.107203.

Wentao Li, Baoxiang Pan, Jiangjiang Xia, and Qingyun Duan. Convolutional neural network-based statistical post-processing of ensemble precipitation forecasts. *Journal of Hydrology*, 605:127301, 2022. https://doi.org/10.1016/j.jhydrol.2021.127301.

Yi Lin and Yongho Jeon. Random forests and adaptive nearest neighbors. *Journal of the American Statistical Association*, 101(474):578–590, 2006. ISSN 1537-274X. https://doi.org/10.1198/016214505000001230.

Edward N. Lorenz. Deterministic nonperiodic flow. *Journal of the Atmospheric Sciences*, 20(2): 130–141, 1963. ISSN 1520-0469. https://doi.org/10.1175/1520-0469(1963)020<0130:dnf>2.0.co;2.

Georges Matheron. Principles of geostatistics. *Economic Geology*, 58(8):1246–1266, 1963. ISSN 0361-0128. https://doi.org/10.2113/gsecongeo.58.8.1246.

James E. Matheson and Robert L. Winkler. Scoring rules for continuous probability distributions. *Management Science*, 22(10):1087–1096, 1976. ISSN 00251909, 15265501. URL http://www.jstor.org/stable/2629907.

Nicolai Meinshausen. Quantile regression forests. *The Journal of Machine Learning Research*, 7(35):983–999, 2006. URL http://jmlr.org/papers/v7/meinshausen06a.html.

Xiaochun Meng, James W. Taylor, Souhaib Ben Taieb, and Siran Li. Scores for multivariate distributions and level sets. *Operations Research*, 2023. ISSN 1526-5463. https://doi.org/10.1287/opre.2020.0365.

Jakob W. Messner, Georg J. Mayr, and Achim Zeileis. Nonhomogeneous boosting for predictor selection in ensemble postprocessing. *Monthly Weather Review*, 145(1):137–147, 2017. https://doi.org/10.1175/mwr-d-16-0088.1.

Allan H. Murphy and Robert L. Winkler. A general framework for forecast verification. *Monthly Weather Review*, 115(7):1330–1338, 1987. ISSN 1520-0493. https://doi.org/10.1175/1520-0493(1987)115<1330:agfffv>2.0.co;2.

Thomas Muschinski, Georg J. Mayr, Achim Zeileis, and Thorsten Simon. Robust weather-adaptive post-processing using model output statistics random forests. *Nonlinear Processes in Geophysics*, 30(4):503–514, 2023. ISSN 1607-7946. https://doi.org/10.5194/npg-30-503-2023.

Annette Möller, Alex Lenkoski, and Thordis L. Thorarinsdottir. Multivariate probabilistic forecasting using ensemble bayesian model averaging and copulas. *Quarterly Journal of the Royal Meteorological Society*, 139(673):982–991, 2013. ISSN 1477-870X. https://doi.org/10.1002/qj.2009.

Alexandre Mösching and Lutz Dümbgen. Monotone least squares and isotonic quantiles. *Electronic Journal of Statistics*, 14(1):24–49, 2020. ISSN 1935-7524. https://doi.org/10.1214/19-ejs1659.

E. A. Nadaraya. On estimating regression. *Theory of Probability & Its Applications*, 9(1): 141–142, 1964. ISSN 1095-7219. https://doi.org/10.1137/1109020.

Philippe Naveau, Raphael Huser, Pierre Ribereau, and Alexis Hannart. Modeling jointly low, moderate, and heavy rainfall intensities without a threshold selection. *Water Resources Research*, 52(4):2753–2769, 2016. https://doi.org/10.1002/2015wr018552.

Jakub Nowotarski and Rafał Weron. Recent advances in electricity price forecasting: A review of probabilistic forecasting. *Renewable and Sustainable Energy Reviews*, 81:1548–1568, 2018. ISSN 1364-0321. https://doi.org/10.1016/j.rser.2017.05.234.

David A. Olson, Norman W. Junker, and Brian Korty. Evaluation of 33 years of quantitative precipitation forecasting at the NMC. *Weather and Forecasting*, 10(3):498–511, 1995. https://doi.org/10.1175/1520-0434(1995)010<0498:eoyoqp>2.0.co;2.

Amila Orucevic, John L. Bell, Megan King, Alison P. McNabb, and Robert E. Heidel. Nomogram update based on tailorx clinical trial results - oncotype dx breast cancer recurrence score can be predicted using clinicopathologic data. *The Breast*, 46:116–125, 2019. ISSN 0960-9776. https://doi.org/10.1016/j.breast.2019.05.006.

Lorenzo Pacchiardi, Rilwan Adewoyin, Peter Dueben, and Ritabrata Dutta. Probabilistic forecasting with generative networks via scoring rule minimization. *Journal of Machine Learning Research*, 25(45):1–64, 2024. URL https://jmlr.org/papers/v25/23-0038.html.

Soonmyung Paik, Steven Shak, Gong Tang, Chungyeul Kim, Joffre Baker, Maureen Cronin, Frederick L. Baehner, Michael G. Walker, Drew Watson, Taesung Park, William Hiller, Edwin R. Fisher, D. Lawrence Wickerham, John Bryant, and Norman Wolmark. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *New England Journal of Medicine*, 351(27):2817–2826, 2004. ISSN 1533-4406. https://doi.org/10.1056/nejmoa041588.

Soonmyung Paik, Gong Tang, Steven Shak, Chungyeul Kim, Joffre Baker, Wanseop Kim, Maureen Cronin, Frederick L. Baehner, Drew Watson, John Bryant, Joseph P. Costantino, Charles E. Geyer, D. Lawrence Wickerham, and Norman Wolmark. Gene expression and benefit of chemotherapy in women with node-negative, estrogen receptor–positive breast cancer. *Journal of Clinical Oncology*, 24(23):3726–3734, 2006. ISSN 1527-7755. https://doi.org/10.1200/jco.2005.04.7985.

T. N. Palmer. The economic value of ensemble forecasts as a tool for risk assessment: From days to decades. *Quarterly Journal of the Royal Meteorological Society*, 128(581):747–774, 2002. ISSN 1477-870X. https://doi.org/10.1256/0035900021643593.

T. N. 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. ISSN 1477-870X. https://doi.org/10.1002/qj.1923.

Victor M. Panaretos and Yoav Zemel. *An Invitation to Statistics in Wasserstein Space*. Springer International Publishing, 2020. ISBN 9783030384388. https://doi.org/10.1007/978-3-030-38438-8.

Matthew Parry, A. Philip Dawid, and Steffen Lauritzen. Proper local scoring rules. *The Annals of Statistics*, 40(1):561 – 592, 2012. ISSN 0090-5364. https://doi.org/10.1214/12-AOS971.

Olivier C. Pasche and Sebastian Engelke. Neural networks for extreme quantile regression with an application to forecasting of flood risk. 2024. https://doi.org/10.48550/ARXIV.2208.07590.

Jaideep Pathak, Shashank Subramanian, Peter Harrington, Sanjeev Raja, Ashesh Chattopadhyay, Morteza Mardani, Thorsten Kurth, David Hall, Zongyi Li, Kamyar Azizzadenesheli, Pedram Hassanzadeh, Karthik Kashinath, and Animashree Anandkumar. Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators. 2022. https://doi.org/10.48550/ARXIV.2202.11214.

Kate R. Pawloski, Mithat Gonen, Hannah Y. Wen, Audree B. Tadros, Donna Thompson, Kelly Abbate, Monica Morrow, and Mahmoud El-Tamer. Supervised machine learning model to predict oncotype dx risk category in patients over age 50. *Breast Cancer Research and Treatment*, 191(2):423–430, 2021. ISSN 1573-7217. https://doi.org/10.1007/s10549-021-06443-w.

Romain Pic, Clément Dombry, Philippe Naveau, and Maxime Taillardat. Distributional regression and its evaluation with the CRPS: Bounds and convergence of the minimax risk. *International Journal of Forecasting*, 39(4):1564–1572, 2023. ISSN 0169-2070. https://doi.org/10.1016/j.ijforecast.2022.11.001.

Romain Pic, Clément Dombry, Philippe Naveau, and Maxime Taillardat. Proper scoring rules for multivariate probabilistic forecasts based on aggregation and transformation. 2024a. https://doi.org/10.48550/ARXIV.2407.00650.

Romain Pic, Clément Dombry, Philippe Naveau, and Maxime Taillardat. Distributional regression u-nets for the postprocessing of precipitation ensemble forecasts. 2024b. https://doi.org/10.48550/ARXIV.2407.02125.

P. Pinson and R. Girard. Evaluating the quality of scenarios of short-term wind power generation. *Applied Energy*, 96:12–20, 2012. https://doi.org/10.1016/j.apenergy.2011.11.004.

Pierre Pinson. Wind energy: Forecasting challenges for its operational management. *Statistical Science*, 28(4):564 – 585, 2013. ISSN 0883-4237. https://doi.org/10.1214/13-STS445.

Pierre Pinson and Julija Tastu. *Discrimination ability of the Energy score*. Number 15 in DTU Compute Technical Report-2013. Technical University of Denmark, 2013.

Pierre Pinson, Henrik Madsen, Henrik Aa. Nielsen, George Papaefthymiou, and Bernd Klöckl. From probabilistic forecasts to statistical scenarios of short-term wind power production. *Wind Energy*, 12(1):51–62, 2009. ISSN 1099-1824. https://doi.org/10.1002/we.284.

Serge Planton, Michel Déqué, Fabrice Chauvin, and Laurent Terray. Expected impacts of climate change on extreme climate events. *Comptes Rendus Geoscience*, 340(9-10):564–574, 2008. https://doi.org/10.1016/j.crte.2008.07.009.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2023. URL https://www.R-project.org/.

Sabine Radanovics, Jean-Philippe Vidal, and Eric Sauquet. Spatial verification of ensemble precipitation: An ensemble version of sal. *Weather and Forecasting*, 33(4):1001–1020, 2018. ISSN 1520-0434. https://doi.org/10.1175/waf-d-17-0162.1.

Stephan Rasp and Sebastian Lerch. Neural networks for postprocessing ensemble weather forecasts. *Monthly Weather Review*, 146(11):3885–3900, 2018. ISSN 1520-0493. https://doi.org/10.1175/mwr-d-18-0187.1.

Stephan Rasp, Stephan Hoyer, Alexander Merose, Ian Langmore, Peter Battaglia, Tyler Russell, Alvaro Sanchez-Gonzalez, Vivian Yang, Rob Carver, Shreya Agrawal, Matthew Chantry, Zied Ben Bouallegue, Peter Dueben, Carla Bromberg, Jared Sisk, Luke Barrington, Aaron Bell,

and Fei Sha. Weatherbench 2: A benchmark for the next generation of data-driven global weather models. *Journal of Advances in Modeling Earth Systems*, 16(6), 2024. ISSN 1942-2466. https://doi.org/10.1029/2023ms004019.

Didier Ricard, Véronique Ducrocq, and Ludovic Auger. A climatology of the mesoscale environment associated with heavily precipitating events over a northwestern mediterranean area. *Journal of Applied Meteorology and Climatology*, 51(3):468–488, 2012. ISSN 1558-8432. https://doi.org/10.1175/jamc-d-11-017.1.

Pauline Rivoire, Olivia Martius, Philippe Naveau, and Alexandre Tuel. Assessment of subseasonal-to-seasonal (s2s) ensemble extreme precipitation forecast skill over europe. *Natural Hazards and Earth System Sciences*, 23(8):2857–2871, 2023. ISSN 1684-9981. https://doi.org/10.5194/nhess-23-2857-2023.

Nigel M. Roberts and Humphrey W. Lean. Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Monthly Weather Review*, 136(1):78–97, 2008. ISSN 0027-0644. https://doi.org/10.1175/2007mwr2123.1.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Lecture Notes in Computer Science*, pages 234–241. Springer International Publishing, 2015. https://doi.org/10.1007/978-3-319-24574-4_28.

Mark S. Roulston and Leonard A. Smith. Evaluating probabilistic forecasts using information theory. *Monthly Weather Review*, 130(6):1653–1660, 2002. ISSN 1520-0493. https://doi.org/10.1175/1520-0493(2002)130<1653:epfuit>2.0.co;2.

Roman Schefzik and Annette Möller. *Ensemble Postprocessing Methods Incorporating Dependence Structures*, pages 91–125. Elsevier, 2018. ISBN 9780128123720. https://doi.org/10.1016/b978-0-12-812372-0.00004-2.

Roman Schefzik, Thordis L. Thorarinsdottir, and Tilmann Gneiting. Uncertainty quantification in complex simulation models using ensemble copula coupling. *Statistical Science*, 28(4):616 – 640, 2013. https://doi.org/10.1214/13-STS443.

Michael Scheuerer and Thomas M. Hamill. Statistical post-processing of ensemble precipitation forecasts by fitting censored, shifted gamma distributions. *Monthly Weather Review*, 143(11): 4578–4596, 2015a. ISSN 1520-0493. https://doi.org/10.1175/MWR-D-15-0061.1.

Michael Scheuerer and Thomas M. Hamill. Variogram-based proper scoring rules for probabilistic forecasts of multivariate quantities∗. *Monthly Weather Review*, 143(4):1321–1334, 2015b. https://doi.org/10.1175/mwr-d-14-00269.1.

Michael Scheuerer, Matthew B. Switanek, Rochelle P. Worsnop, and Thomas M. Hamill. Using artificial neural networks for generating probabilistic subseasonal precipitation forecasts over california. *Monthly Weather Review*, 148(8):3489–3506, 2020. https://doi.org/10.1175/mwr-d-20-0096.1.

Lisa Schlosser, Torsten Hothorn, Reto Stauffer, and Achim Zeileis. Distributional regression forests for probabilistic precipitation forecasting in complex terrain. *The Annals of Applied Statistics*, 13(3):1564 – 1589, 2019. https://doi.org/10.1214/19-AOAS1247.

Danijel Schorlemmer, Maximilian J. Werner, Warner Marzocchi, Thomas H. Jordan, Yosihiko Ogata, David D. Jackson, Sum Mak, David A. Rhoades, Matthew C. Gerstenberger, Naoshi Hirata, Maria Liukis, Philip J. Maechling, Anne Strader, Matteo Taroni, Stefan Wiemer, Jeremy D. Zechar, and Jiancang Zhuang. The collaboratory for the study of earthquake

predictability: Achievements and priorities. *Seismological Research Letters*, 89(4):1305–1313, 2018. ISSN 1938-2057. https://doi.org/10.1785/0220180053.

Benedikt Schulz and Sebastian Lerch. Aggregating distribution forecasts from deep ensembles. 2022a. https://doi.org/10.48550/ARXIV.2204.02291.

Benedikt Schulz and Sebastian Lerch. Machine learning methods for postprocessing ensemble forecasts of wind gusts: A systematic comparison. *Monthly Weather Review*, 150(1):235–257, 2022b. ISSN 1520-0493. https://doi.org/10.1175/mwr-d-21-0150.1.

Erwan Scornet. On the asymptotics of random forests. *Journal of Multivariate Analysis*, 146: 72–83, 2016. ISSN 0047-259X. https://doi.org/10.1016/j.jmva.2015.06.009.

C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(4): 623–656, 1948. ISSN 0005-8580. https://doi.org/10.1002/j.1538-7305.1948.tb00917.x.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. pages 1–14. Computational and Biological Learning Society, 2015. https://doi.org/10.48550/ARXIV.1409.1556.

Alex Smola, Arthur Gretton, Le Song, and Bernhard Schölkopf. A hilbert space embedding for distributions. In Marcus Hutter, Rocco A. Servedio, and Eiji Takimoto, editors, *Algorithmic Learning Theory*, pages 13–31, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg. ISBN 978-3-540-75225-7. https://doi.org/10.1007/978-3-540-75225-7_5.

Joseph A. Sparano, Robert J. Gray, Della F. Makower, Kathleen I. Pritchard, Kathy S. Albain, Daniel F. Hayes, Charles E. Geyer, Elizabeth C. Dees, Edith A. Perez, John A. Olson, JoAnne Zujewski, Tracy Lively, Sunil S. Badve, Thomas J. Saphner, Lynne I. Wagner, Timothy J. Whelan, Matthew J. Ellis, Soonmyung Paik, William C. Wood, Peter Ravdin, Maccon M. Keane, Henry L. Gomez Moreno, Pavan S. Reddy, Timothy F. Goggins, Ingrid A. Mayer, Adam M. Brufsky, Deborah L. Toppmeyer, Virginia G. Kaklamani, James N. Atkins, Jeffrey L. Berenberg, and George W. Sledge. Prospective validation of a 21-gene expression assay in breast cancer. *New England Journal of Medicine*, 373(21):2005–2014, 2015. ISSN 1533-4406. https://doi.org/10.1056/nejmoa1510764.

Joseph A. Sparano, Robert J. Gray, Della F. Makower, Kathleen I. Pritchard, Kathy S. Albain, Daniel F. Hayes, Charles E. Geyer, Elizabeth C. Dees, Matthew P. Goetz, John A. Olson, Tracy Lively, Sunil S. Badve, Thomas J. Saphner, Lynne I. Wagner, Timothy J. Whelan, Matthew J. Ellis, Soonmyung Paik, William C. Wood, Peter M. Ravdin, Maccon M. Keane, Henry L. Gomez Moreno, Pavan S. Reddy, Timothy F. Goggins, Ingrid A. Mayer, Adam M. Brufsky, Deborah L. Toppmeyer, Virginia G. Kaklamani, Jeffrey L. Berenberg, Jeffrey Abrams, and George W. Sledge. Adjuvant chemotherapy guided by a 21-gene expression assay in breast cancer. *New England Journal of Medicine*, 379(2):111–121, 2018. ISSN 1533-4406. https://doi.org/10.1056/nejmoa1804710.

Joël Stein and Fabien Stoop. Neighborhood-based ensemble evaluation using the crps. *Monthly Weather Review*, 150(8):1901–1914, 2022. ISSN 1520-0493. https://doi.org/10.1175/mwr-d-21-0224.1.

Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Information Science and Statistics. Springer, New York, 2008. ISBN 978-0-387-77241-7.

Ingo Steinwart and Johanna F. Ziegel. Strictly proper kernel scores and characteristic kernels on compact spaces. *Applied and Computational Harmonic Analysis*, 51:510–542, 2021. ISSN 1063-5203. https://doi.org/10.1016/j.acha.2019.11.005.

Charles J. Stone. Consistent nonparametric regression. *The Annals of Statistics*, 5(4):595–620, 1977. ISSN 00905364. https://doi.org/10.1214/aos/1176343886.

Charles J. Stone. Optimal rates of convergence for nonparametric estimators. *The Annals of Statistics*, 8(6):1348 – 1360, 1980. https://doi.org/10.1214/aos/1176345206.

Charles J. Stone. Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, 10(4):1040 – 1053, 1982. https://doi.org/10.1214/aos/1176345969.

Maher Sughayer, Rolla Alaaraj, and Ahmad Alsughayer. Applying new magee equations for predicting the oncotype dx recurrence score. *Breast Cancer*, 25(5):597–604, 2018. ISSN 1880-4233. https://doi.org/10.1007/s12282-018-0860-x.

Gábor Székely. E-statistics: The energy of statistical samples. techreport, Bowling Green State University, 2003.

Maxime Taillardat. Skewed and mixture of gaussian distributions for ensemble postprocessing. *Atmosphere*, 12(8):966, 2021. ISSN 2073-4433. https://doi.org/10.3390/atmos12080966.

Maxime Taillardat and Olivier Mestre. From research to applications – examples of operational ensemble post-processing in france using machine learning. *Nonlinear Processes in Geophysics*, 27(2):329–347, 2020. ISSN 1607-7946. https://doi.org/10.5194/npg-27-329-2020.

Maxime Taillardat, Olivier Mestre, Michaël Zamo, and Philippe Naveau. Calibrated ensemble forecasts using quantile regression forests and ensemble model output statistics. *Monthly Weather Review*, 144(6):2375–2393, 2016. ISSN 1520-0493. https://doi.org/10.1175/MWR-D-15-0260.1.

Maxime Taillardat, Anne-Laure Fougères, Philippe Naveau, and Olivier Mestre. Forest-based and semiparametric methods for the postprocessing of rainfall ensemble forecasting. *Weather and Forecasting*, 34(3):617–634, 2019. https://doi.org/10.1175/WAF-D-18-0149.1.

Maxime Taillardat, Anne-Laure Fougères, Philippe Naveau, and Olivier Mestre. Corrigendum. *Weather and Forecasting*, 37(7):1305, 2022. ISSN 1520-0434. https://doi.org/10.1175/waf-d-22-0057.1.

Maxime Taillardat, Anne-Laure Fougères, Philippe Naveau, and Raphaël de Fondeville. Evaluating probabilistic forecasts of extremes using continuous ranked probability score distributions. *International Journal of Forecasting*, 39(3):1448–1459, 2023. ISSN 0169-2070. https://doi.org/10.1016/j.ijforecast.2022.07.003.

O. Talagrand, R. Vautard, and B Strauss. Evaluation of probabilistic prediction systems. In *Workshop on Predictability, 20-22 October 1997*, pages 1–26, Shinfield Park, Reading, 1997. ECMWF.

Thordis L. Thorarinsdottir and Nina Schuhen. *Verification: Assessment of Calibration and Accuracy*, pages 155–186. Elsevier, 2018. https://doi.org/10.1016/b978-0-12-812372-0.00006-6.

Thordis L. Thorarinsdottir, Tilmann Gneiting, and Nadine Gissibl. Using proper divergence functions to evaluate climate models. *SIAM/ASA Journal on Uncertainty Quantification*, 1 (1):522–534, 2013. ISSN 2166-2525. https://doi.org/10.1137/130907550.

J. Thorey, V. Mallet, and P. Baudin. Online learning with the continuous ranked probability score for ensemble forecasting. *Quarterly Journal of the Royal Meteorological Society*, 143 (702):521–529, 2017. ISSN 1477-870X. https://doi.org/10.1002/qj.2940.

Alexandre B. Tsybakov. *Introduction to nonparametric estimation*. Springer Series in Statistics. Springer, New York, 2009. ISBN 978-0-387-79051-0. https://doi.org/10.1007/b13794.

Alexander Tsyplakov. Evaluating density forecasts: A comment. *SSRN Electronic Journal*, 2011. ISSN 1556-5068. https://doi.org/10.2139/ssrn.1907799.

Alexander Tsyplakov. Evaluation of probabilistic forecasts: Proper scoring rules and moments. *SSRN Electronic Journal*, 2013. ISSN 1556-5068. https://doi.org/10.2139/ssrn.2236605.

Alexander Tsyplakov. Evaluation of probabilistic forecasts: Conditional auto-calibration, 2020. URL https://www.sas.upenn.edu/~fdiebold/papers2/Tsyplakov_Auto_calibration_sent_eswc2020.pdf.

Bert Van Schaeybroeck and Stéphane Vannitsem. Ensemble post-processing using member-by-member approaches: theoretical aspects. *Quarterly Journal of the Royal Meteorological Society*, 141(688):807–818, 2015. https://doi.org/10.1002/qj.2397.

Chiem van Straaten, Kirien Whan, and Maurice Schmeits. Statistical postprocessing and multivariate structuring of high-resolution ensemble precipitation forecasts. *Journal of Hydrometeorology*, 19(11):1815–1833, 2018. https://doi.org/10.1175/jhm-d-18-0105.1.

Stéphane Vannitsem, John Bjørnar Bremnes, Jonathan Demaeyer, Gavin R. Evans, Jonathan Flowerdew, Stephan Hemri, Sebastian Lerch, Nigel Roberts, Susanne Theis, Aitor Atencia, Zied Ben Bouallègue, Jonas Bhend, Markus Dabernig, Lesley De Cruz, Leila Hieta, Olivier Mestre, Lionel Moret, Iris Odak Plenković, Maurice Schmeits, Maxime Taillardat, Joris Van den Bergh, Bert Van Schaeybroeck, Kirien Whan, and Jussi Ylhaisi. Statistical postprocessing for weather forecasts: Review, challenges, and avenues in a big data world. *Bulletin of the American Meteorological Society*, 102(3):E681–E699, 2021. ISSN 1520-0477. https://doi.org/10.1175/bams-d-19-0308.1.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. https://doi.org/10.48550/ARXIV.1706.03762.

Simon Veldkamp, Kirien Whan, Sjoerd Dirksen, and Maurice Schmeits. Statistical postprocessing of wind speed forecasts using convolutional neural networks. *Monthly Weather Review*, 149(4):1141–1152, 2021. https://doi.org/10.1175/mwr-d-20-0219.1.

Cédric Villani. *Optimal Transport*. Springer Berlin Heidelberg, 2009. ISBN 9783540710509. https://doi.org/10.1007/978-3-540-71050-9.

Geoffrey S. Watson. Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)*, 26(4):359–372, 1964. ISSN 0581572X. URL http://www.jstor.org/stable/25049340.

Heini Wernli, Marcus Paulat, Martin Hagen, and Christoph Frei. Sal—a novel quality measure for the verification of quantitative precipitation forecasts. *Monthly Weather Review*, 136(11):4470–4487, 2008. ISSN 0027-0644. https://doi.org/10.1175/2008mwr2415.1.

Kirien Whan and Maurice Schmeits. Comparing area probability forecasts of (extreme) local precipitation using parametric and machine learning statistical postprocessing methods. *Monthly Weather Review*, 146(11):3651–3673, 2018. https://doi.org/10.1175/mwr-d-17-0290.1.

Daniel S. Wilks, editor. *Statistical methods in the atmospheric sciences*. Number v. 100 in International Geophysics Series. Academic Press, Oxford, 3rd ed. edition, 2011. ISBN 9780123850232.

Daniel S. Wilks. *Univariate Ensemble Postprocessing*, pages 49–89. Elsevier, 2018. ISBN 9780128123720. https://doi.org/10.1016/b978-0-12-812372-0.00003-0.

Daniel S. Wilks and Stéphane Vannitsem. *Uncertain Forecasts From Deterministic Dynamics*, pages 1–13. Elsevier, 2018. ISBN 9780128123720. https://doi.org/10.1016/b978-0-12-812372-0.00001-7.

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, 2013. https://doi.org/10.1002/qj.2198.

Andreas Winkelbauer. Moments and absolute moments of the normal distribution. 2014. https://doi.org/10.48550/ARXIV.1209.4340.

R. L. Winkler, Javier Muñoz, José L. Cervera, José M. Bernardo, Gail Blattenberger, Joseph B. Kadane, Dennis V. Lindley, Allan H. Murphy, Robert M Oliver, and David Ríos-Insua. Scoring rules and the evaluation of probabilities. *Test*, 5(1):1–60, 1996. ISSN 1863-8260. https://doi.org/10.1007/bf02562681.

Robert L. Winkler. *Rewarding Expertise in Probability Assessment*, pages 127–140. Springer Netherlands, 1977. ISBN 9789401012768. https://doi.org/10.1007/978-94-010-1276-8_10.

Antonio C. Wolff, M. Elizabeth Hale Hammond, Kimberly H. Allison, Brittany E. Harvey, Pamela B. Mangu, John M.S. Bartlett, Michael Bilous, Ian O. Ellis, Patrick Fitzgibbons, Wedad Hanna, Robert B. Jenkins, Michael F. Press, Patricia A. Spears, Gail H. Vance, Giuseppe Viale, Lisa M. McShane, and Mitchell Dowsett. Human epidermal growth factor receptor 2 testing in breast cancer: American society of clinical oncology/college of american pathologists clinical practice guideline focused update. *Journal of Clinical Oncology*, 36(20): 2105–2122, 2018. ISSN 1527-7755. https://doi.org/10.1200/jco.2018.77.8738.

Marvin N. Wright and Andreas Ziegler. ranger: A fast implementation of random forests for high dimensional data in c++ and r. *Journal of Statistical Software*, 77(1):1–17, 2017. ISSN 1548-7660. https://doi.org/10.18637/jss.v077.i01.

B Yeo, L Zabaglo, M Hills, A Dodson, I Smith, and M Dowsett. Clinical utility of the ihc4+c score in oestrogen receptor-positive early breast cancer: a prospective decision impact study. *British Journal of Cancer*, 113(3):390–395, 2015. ISSN 1532-1827. https://doi.org/10.1038/bjc.2015.222.

Michaël Zamo. *Statistical Post-processing of Deterministic and Ensemble Wind Speed Forecasts on a Grid*. Theses, Université Paris Saclay (COmUE), 2016. URL https://theses.hal.science/tel-01598119.

Michaël Zamo and Philippe Naveau. Estimation of the continuous ranked probability score with limited information and applications to ensemble weather forecasts. *Mathematical Geosciences*, 50(2):209–234, 2017. ISSN 1874-8953. https://doi.org/10.1007/s11004-017-9709-7.

Alaa Zare, Lynne-Marie Postovit, and John Maringa Githaka. Robust inflammatory breast cancer gene signature using nonparametric random forest analysis. *Breast Cancer Research*, 23(1):92, 2021. ISSN 1465-542X. https://doi.org/10.1186/s13058-021-01467-y.

Florian Ziel and Kevin Berk. Multivariate forecasting evaluation: On sensitive and strictly proper scoring rules. 2019. https://doi.org/10.48550/ARXIV.1910.07325.

Domagoj Ćevid, Loris Michel, Jeffrey Näf, Nicolai Meinshausen, and Peter Bühlmann. Distributional random forests: Heterogeneity adjustment and multivariate distributional regression. *Journal of Machine Learning Research*, 23(333):1–79, 2022. URL http://jmlr.org/papers/v23/21-0585.html.

Appendix

Appendix A

Stone's theorem for distributional regression in Wasserstein distance

This chapter reproduces an article published in *Journal of Nonparametric Statistics* and written by Clément Dombry[1], Thibault Modeste[2] and Romain Pic[1].

Abstract We extend the celebrated Stone's theorem to the framework of distributional regression. More precisely, we prove that weighted empirical distributions with local probability weights satisfying the conditions of Stone's theorem provide universally consistent estimates of the conditional distributions, where the error is measured by the Wasserstein distance of order $p \geq 1$. Furthermore, for $p = 1$, we determine the minimax rates of convergence on specific classes of distributions. We finally provide some applications of these results, including the estimation of conditional tail expectation or probability-weighted moments.

Contents

A.1 Introduction

Forecasting is a major task from statistics and is often of crucial importance for decision-making. In the simple case when the quantity of interest is univariate and quantitative, point forecasting

[1]Université de Franche Comté, CNRS, LmB (UMR 6623), F-25000 Besançon, France
[2]Université Claude Bernard Lyon 1, CNRS, UMR 5208, Institut Camille Jordan, 69622 Villeurbanne, France

often takes the form of regression where one aims at estimating the conditional mean (or the conditional quantile) of the response variable Y given the available information encoded in a vector of covariates X. A point forecast is only a rough summary statistic and should at least be accompanied by an assessment of uncertainty (e.g. standard deviation or a confidence interval). Alternatively, probabilistic forecasting and distributional regression (Gneiting and Katzfuss, 2014) suggest estimating the full conditional distribution of Y given X, called the predictive distribution.

In the last decades, weather forecasting has been a major motivation for the development of probabilistic forecasts. Ensemble forecasts are based on a given number of deterministic models whose parameters vary slightly in order to take into account observation errors and incomplete physical representation of the atmosphere. This leads to an ensemble of different forecasts that overall also assess the uncertainty of the forecast. Ensemble forecasts suffer from bias and underdispersion (Hamill and Colucci, 1997) and need to be statistically postprocessed in order to be improved. Different postprocessing methods have been proposed, such as Ensemble Model Output Statistics (Gneiting et al., 2005), Quantile Regression Forests (Taillardat et al., 2019) or Neural Networks (Schulz and Lerch, 2022b) among others. Distributional regression is now widely used beyond meteorology and recent methodological works include deep distribution regression by Li et al. (2021), distributional random forest by Ćevid et al. (2022) or isotonic distributional regression by Henzi et al. (2021).

The purpose of the present paper is to provide an extension to the framework of distributional regression of the celebrated Stone's theorem (Stone, 1977) that states the consistency of the local weight algorithm for the estimation of the regression function. The strength of Stone's theorem is that it is fully non-parametric and model-free, with very mild assumptions that cover many important cases such as kernel algorithms and nearest neighbor methods, see e.g. Györfi et al. (2002) for more details. We prove that Stone's theorem has a natural and elegant extension to distributional regression with error measured by the Wasserstein distance of order $p \geq 1$. Our result covers not only the case of a one-dimensional output $Y \in \mathbb{R}$ where the Wasserstein distance has a simple explicit form, but also the case of a multivariate output $Y \in \mathbb{R}^d$. The use of the Wasserstein distance is motivated by recent works revealing that it is a useful and powerful tool in statistics, see, e.g., the review by Panaretos and Zemel (2020). Besides this main result, we characterize, in the case $d = 1$ and $p = 1$, the optimal minimax rate of convergence on suitable classes of distributions. We also discuss implications of our results to estimate various statistics of possible interest such as the expected shortfall or the probability-weighted moment.

The structure of the paper is the following. In Section A.2, we present the required background on Stone's theorem and Wasserstein spaces. Section A.3 gathers our main results, including the extension of Stone's theorem to distributional regression (Theorem A.2), the characterization of optimal minimax rates of convergence (Theorem A.3) and some applications (Proposition A.2 and the subsequent examples). All the technical proofs are postponed to Section A.4.

A.2 Background

A.2.1 Stone's theorem

In a regression framework, we observe a sample (X_i, Y_i), $1 \leq i \leq n$, of independent copies of $(X, Y) \in \mathbb{R}^k \times \mathbb{R}^d$ with distribution P. Based on this sample and assuming Y integrable, the goal is to estimate the regression function

$$r(x) = \mathbb{E}[Y|X = x], \quad x \in \mathbb{R}^k.$$

Local average estimators take the form

$$\hat{r}_n(x) = \sum_{i=1}^{n} W_{ni}(x) Y_i \tag{A.1}$$

with $W_{n1}(x), \ldots, W_{nn}(x)$ the *local weights* at x. The local weights are assumed to be measurable functions of x and $X_1, \ldots, X_n$ but not to depend on $Y_1, \ldots, Y_n$, that is

$$W_{ni}(x) = W_{ni}(x; X_1, \ldots, X_n), \quad 1 \leq i \leq n. \tag{A.2}$$

For the convenience of notation, the dependency on $X_1, \ldots, X_n$ is implicit. In this paper, we focus only on the case of *probability weights* satisfying

$$W_{ni}(x) \geq 0, \ 1 \leq i \leq n, \quad \text{and} \quad \sum_{i=1}^{n} W_{ni}(x) = 1. \tag{A.3}$$

Stone's Theorem states the universal consistency of the regression estimate in L^p-norm.

Theorem A.1 (Stone 1977)**.** *Assume the probability weights* (A.3) *satisfy the following three conditions:*

i) there is $C > 0$ *such that* $\mathbb{E}\left[\sum_{i=1}^{n} W_{ni}(X) g(X_i)\right] \leq C\mathbb{E}[g(X)]$ *for all* $n \geq 1$ *and measurable* $g : \mathbb{R}^k \to [0, +\infty)$ *such that* $\mathbb{E}[g(X)] < \infty$*;*

ii) for all $\varepsilon > 0$, $\sum_{i=1}^{n} W_{ni}(X) \mathbb{1}_{\{\|X_i - X\| > \varepsilon\}} \to 0$ *in probability as* $n \to +\infty$*;*

iii) $\max_{1 \leq i \leq n} W_{ni}(X) \to 0$ *in probability as* $n \to +\infty$*.*

Then, for all $p \geq 1$ *and* $(X, Y) \sim P$ *such that* $\mathbb{E}[\|Y\|^p] < \infty$*,*

$$\mathbb{E}\left[\|\hat{r}_n(X) - r(X)\|^p\right] \longrightarrow 0 \quad \text{as } n \to +\infty. \tag{A.4}$$

Conversely, if Equation (A.4) *holds, then the probability weights must satisfy conditions* $i) - iii)$*.*

Remark A.1. Stone's theorem is usually stated in dimension $d = 1$. Since the convergence of random vectors $\hat{r}_n(X) \to r(X)$ in L^p is equivalent to convergence in L^p of all the components, the extension to the dimension $d \geq 2$ is straightforward. Furthermore, more general weights than probability weights can be considered: condition (A.3) can be dropped and replaced by the weaker assumptions that

$$|W_{ni}(X)| \leq M \quad \text{a.s. for some } M > 0.$$

and

$$\sum_{i=1}^{n} W_{ni}(X) \to 1 \text{ in probability.}$$

Such general weights will not be considered in the present paper and we therefore stick to probability weights. The reader can refer to Biau and Devroye (2015) for a complete proof of Stone's theorem together with a discussion.

Example A.1. The following two examples of kernel weights and nearest neighbor weights are the most important ones in the literature and we refer to Györfi et al. (2002) Chapter 5 and 6, respectively, for more details.

- The kernel weights are defined by

$$W_{ni}(x) = \frac{K\left(\frac{x-X_i}{h_n}\right)}{\sum_{j=1}^{n} K\left(\frac{x-X_j}{h_n}\right)}, \quad 1 \leq i \leq n \tag{A.5}$$

if the denominator is nonzero, and $1/n$ otherwise. Here the bandwidth $h_n > 0$ depends only on the sample size n and the function $K : \mathbb{R}^k \to [0, +\infty)$ is called a kernel. In this case, the estimator (A.1) corresponds to the Nadaraya-Watson estimator of the regression function (Nadaraya, 1964; Watson, 1964). We say that K is a boxed kernel if there are constants $R_2 \geq R_1 > 0$ and $M_2 \geq M_1 > 0$ such that

$$M_1 \mathbb{1}_{\{\|x\| \leq R_1\}} \leq K(x) \leq M_2 \mathbb{1}_{\{\|x\| \leq R_2\}}, \quad x \in \mathbb{R}^k.$$

Theorem 5.1 in Györfi et al. (2002) states that, for a boxed kernel, the kernel weights (A.5) satisfy conditions $i) - iii)$ of Theorem A.1 if and only if $h_n \to 0$ and $nh_n^k \to +\infty$ as $n \to +\infty$.

- The nearest neighbor (NN) weights are defined by

$$W_{ni}(x) = \begin{cases} \frac{1}{\kappa_n} & \text{if } X_i \text{ belongs to the } \kappa_n\text{-NN of } x \\ 0 & \text{otherwise} \end{cases}, \tag{A.6}$$

where the number of neighbors $\kappa_n \in \{1, \ldots, n\}$ depends only on the sample size. Recall that the κ_n-NN of x within the sample $(X_i)_{1 \leq i \leq n}$ are obtained by sorting the distances $\|X_i - x\|$ in increasing order and keeping the κ_n points with the smallest distances – as discussed in Györfi et al. (2002) Chapter 6, several rules can be used to break ties such as lexicographic or random tie-breaking. Theorem 6.1 in the same reference states that the nearest neighbor weights (A.6) satisfy conditions $i) - iii)$ of Theorem A.1 if and only if $\kappa_n \to +\infty$ and $\kappa_n/n \to 0$ as $n \to +\infty$.

Example A.2. Interestingly, some variants of the celebrated Breiman's Random Forest (Breiman, 2001) produce probability weights satisfying the assumptions of Stone's theorem. In Breiman's Random Forest, the splits involve both the covariates and the response variable so that the associated weights $W_{ni}(x) = W_{ni}(x; (X_l, Y_l)_{1 \leq l \leq n})$ are not in the form (A.2). Scornet (2016) considers two simplified versions of infinite random forests where the associated weights $W_{ni}(x)$ do not depend on the response values and satisfy the so-called X-property, that is they are in the form (A.2). For totally non-adaptive forests, the trees are grown thanks to a binary splitting rule that does not use the training sample and is totally random; the author shows that the probability weights associated to the infinite forest satisfy the assumptions of Stone's theorem under the condition that the number of leaves grows to infinity at a rate smaller than n and the leaf volume tends to zero in probability (see Theorem 4.1 and its proof). For q-quantile forests, the binary splitting rule involves only the covariates and the author shows that the weights associated to the infinite forest satisfy the assumptions of Stone's theorem provided the subsampling number a_n satisfies $a_n \to +\infty$ and $a_n/n \to 0$ (see Theorem 5.1 and its proof).

A.2.2 Wasserstein spaces

We recall the definition and some elementary facts on Wasserstein spaces on $\mathbb{R}^d$. More details and further results on optimal transport and Wasserstein spaces can be found in the monograph by Villani (2009), Chapter 6.

For $p \geq 1$, the Wasserstein space $\mathcal{W}_p(\mathbb{R}^d)$ is defined as the set of Borel probability measures on $\mathbb{R}^d$ having a finite moment of order p, i.e. such that

$$M_p(\mu) = \Big(\int_{\mathbb{R}^d} \|y\|^p \, \mu(\mathrm{d}y) \Big)^{1/p} < \infty. \tag{A.7}$$

It is endowed with the distance defined, for $Q_1, Q_2 \in \mathcal{W}_p(\mathbb{R}^d)$, by

$$\mathcal{W}_p(Q_1, Q_2) = \inf_{\pi \in \Pi(Q_1, Q_2)} \left(\int \|y_1 - y_2\|^p \, \pi(\mathrm{d}y_1 \mathrm{d}y_2) \right)^{1/p}, \tag{A.8}$$

where $\Pi(Q_1, Q_2)$ denotes the set of measures on $\mathbb{R}^d \times \mathbb{R}^d$ with marginal distributions Q_1 and Q_2. A couple (Z_1, Z_2) of random variables with distributions Q_1 and Q_2 respectively is called a *coupling*. The Wasserstein distance is thus the minimal distance $\|Z_1 - Z_2\|_{\mathrm{L}^p} = \mathbb{E}[\|Z_1 - Z_2\|^p]^{1/p}$ over all possible couplings. Existence of optimal couplings is ensured since $\mathbb{R}^d$ is a complete and separable metric space so that the infimum is indeed a minimum.

Wasserstein distances are generally difficult to compute, but the case $d = 1$ is the exception. A simple optimal coupling is provided by the probability inverse transform: for $i = 1, 2$, let $Q_i \in \mathcal{W}_p(\mathbb{R})$, F_i denotes its cumulative distribution function and F_i^{-1} its generalized inverse (quantile function). Then, starting from a uniform random variable $U \sim \mathrm{Unif}(0, 1)$, an optimal coupling is given by $(Z_1, Z_2) = (F_1^{-1}(U), F_2^{-1}(U))$. Therefore, the Wasserstein distance is explicitly given by

$$\mathcal{W}_p(Q_1, Q_2) = \left(\int_0^1 |F_1^{-1}(u) - F_2^{-1}(u)|^p \mathrm{d}u \right)^{1/p}. \tag{A.9}$$

When $p = 1$, we have the simpler formula

$$\mathcal{W}_1(Q_1, Q_2) = \int_{-\infty}^{+\infty} |F_1(x) - F_2(x)| \mathrm{d}x. \tag{A.10}$$

which follows from the computation

$$\begin{aligned}
\int_0^1 |F_1^{-1}(u) - F_2^{-1}(u)| \mathrm{d}u &= \int_0^1 \int_{-\infty}^{+\infty} |\mathbb{1}_{x \leq F_1^{-1}(u)} - \mathbb{1}_{x \leq F_1^{-1}(u)}| \mathrm{d}x \mathrm{d}u \\
&= \int_{-\infty}^{+\infty} \int_0^1 |\mathbb{1}_{F_1(x) \leq u} - \mathbb{1}_{F_2(x) \leq u}| \mathrm{d}u \mathrm{d}x \\
&= \int_{-\infty}^{+\infty} |F_1(x) - F_2(x)| \mathrm{d}x.
\end{aligned}$$

A.3 Main results

A.3.1 Stone's theorem for distributional regression

We now present the main result of the paper which is a natural extension of Stone's theorem to the framework of distributional regression. Given a distribution $(X, Y) \sim P$ on $\mathbb{R}^k \times \mathbb{R}^d$, we denote by F the marginal distribution of Y and by F_x its conditional distribution given $X = x$. This conditional distribution can be estimated on a sample $(X_i, Y_i)_{1 \leq i \leq n}$ of independent copies of (X, Y) by the weighted empirical distribution

$$\hat{F}_{n,x} = \sum_{i=1}^n W_{ni}(x) \delta_{Y_i} \tag{A.11}$$

where δ_y denotes the Dirac mass at point $y \in \mathbb{R}^d$. For probability weights satisfying (A.3), $\hat{F}_{n,x}$ is a probability measure and can be viewed as a random element in the complete and separable space $\mathcal{W}_p(\mathbb{R}^d)$. We recall that the weights $W_{ni}(x) = W_{ni}(x; X_1, \ldots, X_n)$ implicitly depend on $X_1, \ldots, X_n$ but not on $Y_1, \ldots, Y_n$.

Theorem A.2. *Assume the probability weights satisfy conditions i) − iii) from Theorem A.1. Then, for all $p \geq 1$ and $(X, Y) \sim P$ such that $\mathbb{E}[\|Y\|^p] < \infty$,*

$$\mathbb{E}\big[\mathcal{W}_p^p(\hat{F}_{n,X}, F_X)\big] \longrightarrow 0 \quad \text{as } n \to +\infty. \tag{A.12}$$

Conversely, if Equation (A.12) *holds, then the probability weights must satisfy conditions i)−iii).*

It is worth noticing that

$$\mathbb{E}\left[\|\hat{r}_n(X) - r(X)\|^p\right] \leq \mathbb{E}\big[\mathcal{W}_p^p(\hat{F}_{n,X}, F_X)\big]$$

so that Theorem A.2 implies Theorem A.1 in a straightforward way. The proof of Theorem A.2 is postponed to Section A.4. It first considers the case $d = 1$ where the Wasserstein distance is explicitly given by formula (A.9). Then, the result is extended to higher dimension $d \geq 2$ thanks to the notion of max-sliced Wasserstein distance (Bayraktar and Guo, 2021) which allows to reduce the convergence of measures on $\mathbb{R}^d$ to the convergence of their one-dimensional projections (a precise statement is given in Theorem A.4 below).

Remark A.2. The fact that the covariate X takes its values in $\mathbb{R}^k$ is not necessary and Theorem A.2 holds for covariate in an abstract metric space. Assuming that $(X, Y) \in \mathcal{X} \times \mathbb{R}^d$ with $\mathcal{X}$ a metric space, it is straightforward to check that the proof of Theorem A.2 goes through with no further complication. This can be useful for instance in the framework of statistical postprocessing of probabilistic forecast where the covariate itself is a probability distribution, see, e.g., Gneiting and Katzfuss (2014). This remark has been suggested by an anonymous referee whom we wish to thank.

A.3.2 Rates of convergence

We next consider rates of convergence in the minimax sense. Note that similar questions and results have been established in Pic et al. (2023), where the second-order Cramér's distance was considered, i.e.

$$\|\hat{F}_{n,X} - F_X\|_{L_2}^2 = \int_{\mathbb{R}} |\hat{F}_{n,X}(y) - F_X(y)|^2 \, \mathrm{d}y.$$

We focus here on the Wasserstein distance $\mathcal{W}_p(\hat{F}_{n,X}, F_X)$ and consider only the case $d = 1$ and $p = 1$ which allows the explicit expression (A.10). The other cases seem harder to analyze and are beyond the scope of the present paper. Our first result considers the error in Wasserstein distance when $X = x$ is fixed.

Proposition A.1. *Assume $d = 1$ and $(X, Y) \sim P$ such that $\mathbb{E}[|Y|] < \infty$. Then,*

$$\mathbb{E}\big[\mathcal{W}_1(\hat{F}_{n,x}, F_x)\big] \leq \mathbb{E}\Big[\sum_{i=1}^n W_{ni}(x)\mathcal{W}_1(F_{X_i}, F_x)\Big] + M(x)\mathbb{E}\Big[\sum_{i=1}^n W_{ni}^2(x)\Big]^{1/2},$$

where $M(x) = \int_{\mathbb{R}} \sqrt{F_x(z)(1 - F_x(z))}\mathrm{d}z$.

The first term corresponds to an approximation error due to the fact that we use a biased sample to estimate F_x. The more regular the model is, the smaller the approximation error is. The second term is an estimation error due to the fact that we use an empirical mean to estimate F_x. This estimator error is smaller if the distribution error has a lower dispersion (as measured by $M(x)$) or if $\sum_{i=1}^n W_{ni}^2(x)$ is small. Note that in the case of nearest neighbor weights, $1/\sum_{i=1}^n W_{ni}^2(x)$ is exactly equal to κ so that this quantity is often referred to as the *effective sample size* and the estimation error is proportional to the square root of the expected reciprocal effective sample size.

In view of Proposition A.1, we introduce the following classes of functions.

Definition A.1. *Let $\mathcal{D}(H, L, M)$ be the class of distributions $(X, Y) \sim P$ on $\mathbb{R}^k \times \mathbb{R}$ satisfying:*

a) $X \in [0, 1]^k$ a.s. and $\mathbb{E}|Y| < \infty$,

b) for all $x, x' \in [0, 1]^k$, $\mathcal{W}_1(F_x, F_{x'}) \leq L\|x - x'\|^H$,

c) for all $x \in [0, 1]^k$, $\int_{\mathbb{R}} \sqrt{F_x(z)(1 - F_x(z))}\, \mathrm{d}z \leq M$.

The definition of the class together with Proposition A.1 entails that the expected error is uniformly bounded on the class $\mathcal{D}(H, L, M)$ by

$$\begin{aligned}&\mathbb{E}\Big[\mathcal{W}_1(\hat{F}_{n,X}, F_X)\Big]\\&\leq L\mathbb{E}\Big[\sum_{i=1}^n W_{ni}(X)\|X_i - X\|^H\Big] + M\mathbb{E}\Big[\sum_{i=1}^n W_{ni}^2(X)\Big]^{1/2}.\end{aligned} \tag{A.13}$$

As a consequence, Proposition A.1 allows to derive explicit bounds uniformly on $\mathcal{D}(H, L, M)$ for the kernel and nearest neighbor methods from Example A.1. For the sake of simplicity, we consider the uniform kernel only.

Corollary A.1. *Let $\hat{F}_{n,X}$ be given by the kernel method with uniform kernel $K(x) = \mathbb{1}_{\{\|x\|\leq 1\}}$ and weights given by Equation* (A.5)*. If $P \in \mathcal{D}(H, L, M)$, then*

$$\mathbb{E}\big[\mathcal{W}_1(\hat{F}_{n,X}, F_X)\big] \leq Lh_n^H + M\sqrt{(2+1/n)c_k}(nh_n^k)^{-1/2} + Lk^{H/2}c_k(nh_n^k)^{-1}$$

with $c_k = k^{k/2}$.

Corollary A.2. *Let $\hat{F}_{n,X}$ be given by the nearest neighbor method with weights given by Equation* (A.6) *and assume $P \in \mathcal{D}(H, L, M)$. Then,*

$$\mathbb{E}\big[\mathcal{W}_1(\hat{F}_{n,X}, F_X)\big] \leq \begin{cases} L8^{H/2}(\kappa_n/n)^{H/2} + M\kappa_n^{-1/2} & \text{if } k = 1,\\ L\tilde{c}_k^{H/2}(\kappa_n/n)^{H/k} + M\kappa_n^{-1/2} & \text{if } k \geq 2,\end{cases}$$

where $\tilde{c}_k$ depends only on the dimension k and is defined in Biau and Devroye (2015, Theorem 2.4).

One can see that consistency holds — i.e. the expected error tends to 0 as $n \to +\infty$ — as soon as $h_n \to 0$ and $nh_n^k \to +\infty$ for the kernel method and $\kappa_n/n \to 0$ and $\kappa_n \to +\infty$ for the nearest neighbor method.

The next theorem provides the optimal minimax rate of convergence on the class $\mathcal{D}(H, L, M)$. We say that two sequences of positive numbers (a_n) and (b_n) have the same rate of convergence, noted $a_n \asymp b_n$, if the ratios a_n/b_n and b_n/a_n remain bounded as $n \to +\infty$.

Theorem A.3. *The optimal minimax rate of convergence on the class $\mathcal{D}(H, L, M)$ is given by*

$$\inf_{\hat{F}_n} \sup_{P\in\mathcal{D}(H,L,M)} \mathbb{E}[\mathcal{W}_1(\hat{F}_{n,X}, F_X)] \asymp n^{-H/(2H+k)}.$$

Theorem A.3 is the counterpart of Pic et al. (2023, Theorem 1) where the minimax rate of convergence for the second order Cramér's distance has been considered. The strategy of proof is similar: i) we prove a lower bound by considering a suitable class of binary distributions where the error in Wasserstein distance corresponds to an absolute error in point regression for which the minimax lower rate of convergence is known; ii) we check that the upper bound for the kernel and/or nearest neighbor algorithm has the same rate of convergence as the lower bound, which proves that the optimal minimax rate of convergence has been identified. In particular, our proof shows that the kernel method defined in Equation (A.5) reaches the minimax rate of convergence in any dimension $k \geq 1$ with the choice of bandwidth $h_n \asymp n^{-1/(2H+k)}$; the nearest neighbor method defined in Equation (A.6) reaches the minimax rate of convergence in any dimension $k \geq 2$ with the number of neighbors $\kappa_n \asymp n^{H/(H+k/2)}$.

Remark A.3. Our estimate of the minimax rate of convergence holds only for $d = p = 1$ and we briefly discuss what can be expected in other cases.
When $p = 1$ and $d \geq 2$, one may consider the Kantorovitch duality expressing the Wasserstein

distance as an integral probability metric. More precisely, in Proposition A.1 and its proof, one would need to study the estimation error

$$\mathcal{W}_1(\tilde{F}_{n,x}, F_x) = \sup_{\mathrm{Lip}(f)\geq 1} \int_{\mathbb{R}^d} f(y)(\tilde{F}_{n,x}, F_x)(\mathrm{d}y)$$

where the supremum is taken over Lipschitz continuous functions with Lipschitz constant smaller than 1. The properties of the weighted empirical process $\sum_{i=1}^n W_{ni}(x)(\delta_{\tilde{Y}_i} - F_x)$ should be useful.

When $p > 1$, even in dimension $d = 1$, it seems much more difficult to obtain bounds for the Wasserstein distance of order p because no simple expression is available. For empirical distributions (without weights), the theory is already involved: a strong result due to Fournier and Guillin (2015) is that, provided $p > d/2$ and integrability condition on Y of order $q > 2p$, the rate of convergence of the empirical distribution $\hat{F}_n = \frac{1}{n}\sum_{i=1}^n \delta_{Y_i}$ to F in $W_p(\mathbb{R}^d)$ is of order $O(\sqrt{n})$ for an i.i.d. sample $Y_1, \ldots, Y_n$ with distribution F on $\mathbb{R}$. One strategy could be to consider the extension of such a result to weighted empirical distributions.

A.3.3 Applications

We briefly illustrate Theorem A.2 with some applications and examples. In statistics, we commonly face the following generic situation: we are interested in a summary statistic S with real values, e.g. quantiles or tail expectation, and we want to assess the effect of X on Y through S, that is we want to assess $S_{Y|X=x}$. Assuming that S is well-defined for distributions on $\mathbb{R}^d$ with a finite moment of order $p \geq 1$, it can be seen as a map $S : \mathcal{W}_p(\mathbb{R}^d) \to \mathbb{R}$ and then $S_{Y|X=x} = S(F_x)$ with F_x the conditional distribution of Y given $X = x$. A natural plug-in estimate of $S_{Y|X=x}$ is

$$\hat{S}_{n,x} = S(\hat{F}_{n,x}) \quad \text{with } \hat{F}_{n,x} \text{ defined by (A.11).}$$

In this generic situation, our extension of Stone's theorem directly implies the following proposition. Recall that $M_p(\mu)$ is defined in Equation (A.7).

Proposition A.2. *Assume $\mathbb{E}[\|Y\|^p] < \infty$ and $\mathbb{P}(F_X \in \mathcal{C}) = 1$ where $\mathcal{C} \subset \mathcal{W}_p(\mathbb{R}^d)$ denotes the continuity set of the statistic $S : \mathcal{W}_p(\mathbb{R}^d) \to \mathbb{R}$. Then weak consistency holds, i.e.*

$$\hat{S}_{n,X} \longrightarrow S_{Y|X} \quad \textit{in probability as } n \to +\infty.$$

If furthermore the statistic S admits a bound of the form

$$|S(\mu)| \leq a M_p^q(\mu) + b, \quad \textit{with } a, b \geq 0 \textit{ and } 0 < q \leq p, \tag{A.14}$$

then consistency holds in $\mathrm{L}^{p/q}$, i.e.

$$\mathbb{E}\big[|\hat{S}_{n,X} - S_{Y|X}|^{p/q}\big] \longrightarrow 0 \quad \textit{as } n \to +\infty$$

Example A.3. (quantile). For a distribution G on $\mathbb{R}$, we define the associated quantile function

$$G^{-1}(\alpha) = \inf\{z \in \mathbb{R} : G(z) \geq \alpha\}, \quad 0 < \alpha < 1.$$

It is well-known that the weak convergence $G_n \xrightarrow{d} G$ implies the quantile convergence $G_n^{-1}(\alpha) \to G^{-1}(\alpha)$ at each continuity point α of G^{-1}. Equivalently, considering $\mathcal{P}(\mathbb{R})$ endowed with the weak convergence topology, the α-quantile statistic $S_\alpha(G) = G^{-1}(\alpha)$ is continuous at G as soon as G^{-1} is continuous at α.

In view of this, we let $\mathcal{C} = \{G \in \mathcal{P}(\mathbb{R})\colon G^{-1} \text{ continuous on } (0,1)\}$ and assume that the conditional distribution satisfies $\mathbb{P}(F_X \in \mathcal{C}) = 1$. Then weak convergence holds for the conditional quantiles, i.e.

$$\hat{F}_{n,X}^{-1}(\alpha) \to F_X^{-1}(\alpha) \quad \text{in probability.}$$

Note that no integrability condition is needed here because we can apply Proposition A.2 on the transformed data $(X_i, \tilde{Y}_i)_{1\leq i\leq n}$, where $\tilde{Y}_i = \tan^{-1}(Y_i)$ is bounded so that convergence in Wasserstein distance is equivalent to weak convergence. If furthermore Y is p-integrable, then the bound

$$\begin{aligned}|S_\alpha(G)|^p &\leq \frac{1}{\alpha}\int_0^\alpha |G^{-1}(u)|^p \mathrm{d}u + \frac{1}{1-\alpha}\int_\alpha^1 |G^{-1}(u)|^p \mathrm{d}u \\ &\leq \Big(\frac{1}{\alpha} + \frac{1}{1-\alpha}\Big) M_p^p(G)\end{aligned}$$

implies the strengthened convergence

$$\hat{F}_{n,X}^{-1}(\alpha) \to F_X^{-1}(\alpha) \quad \text{in } \mathrm{L}^p.$$

Example A.4. (tail expectation) The tail expectation above level $\alpha \in (0,1)$ is the risk measure defined for $G \in \mathcal{W}_1(\mathbb{R})$ by

$$S_\alpha(G) = \frac{1}{1-\alpha}\int_\alpha^1 G^{-1}(u)\,\mathrm{d}u.$$

The name comes from the equivalent definition

$$S_\alpha(G) = \mathbb{E}[Y \mid Y > G^{-1}(\alpha)], \quad Y \sim G,$$

which holds when G^{-1} is continuous at α. One can see that

$$\begin{aligned}|S_\alpha(G_1) - S_\alpha(G_2)| &\leq \frac{1}{1-\alpha}\int_\alpha^1 |G_1^{-1}(u) - G_1^{-1}(u)|\,\mathrm{d}u \\ &\leq \frac{1}{1-\alpha}\int_0^1 |G_1^{-1}(u) - G_2^{-1}(u)|\,\mathrm{d}u \\ &= \frac{1}{1-\alpha}\mathcal{W}_1(G_1, G_2).\end{aligned}$$

so that S_α is Lipschitz continuous with respect to the Wasserstein distance $\mathcal{W}_1$. As a consequence, the conditional tail expectation $S_\alpha(F_x)$ can be estimated in a consistent way by the plug-in estimator $S_\alpha(\hat{F}_{n,x})$ since

$$\mathbb{E}[|S_\alpha(\hat{F}_{n,X}) - S_\alpha(F_X)|] \leq \frac{1}{1-\alpha}\mathbb{E}[\mathcal{W}_1(\hat{F}_{n,X}, F_X)] \longrightarrow 0.$$

Example A.5. (probability-weighted moments, Greenwood et al. 1979) A similar result holds for the probability-weighted moment of order $p, q > 0$ defined by

$$S_{p,q}(G) = \int_0^1 G^{-1}(u)u^p(1-u)^q\,\mathrm{d}u, \quad G \in \mathcal{W}_1(\mathbb{R}).$$

The name comes from the equivalent definition

$$S(G) = \mathbb{E}[YG(Y)^p(1-G(Y))^q], \quad Y \sim G,$$

which holds when G^{-1} is continuous on $(0,1)$. One can again check that the statistic $S_{p,q}$ is Lipschitz continuous with respect to the Wasserstein distance $\mathcal{W}_1$ since

$$\begin{aligned}|S_{p,q}(G_1) - S_{p,q}(G_2)| &\leq \int_0^1 |G_1^{-1}(u) - G_2^{-1}(u)|u^p(1-u)^q\,\mathrm{d}u \\ &\leq \max_{0\leq u\leq 1} u^p(1-u)^q \times \int_0^1 |G_1^{-1}(u) - G_2^{-1}(u)|\,\mathrm{d}u \\ &= \Big(\frac{p}{p+q}\Big)^p\Big(\frac{q}{p+q}\Big)^q \mathcal{W}_1(G_1, G_2).\end{aligned}$$

Example A.6. (covariance) We conclude with a simple example in dimension $d = 2$ where the statistic of interest is the covariance between the two components of $Y = (Y_1, Y_2)$ given $X = x$. Here, we consider

$$S(G) = \int_{\mathbb{R}^2} y_1 y_2 \, \mathrm{d}G - \int_{\mathbb{R}^2} y_1 \, \mathrm{d}G \int_{\mathbb{R}^2} y_2 \, \mathrm{d}G, \quad G \in \mathcal{W}_2(\mathbb{R}^2).$$

Considering square-integrable random vectors $Y = (Y_1, Y_2)$ and $Z = (Z_1, Z_2)$ with distribution G and H respectively, we compute

$$\begin{aligned}
&|S(G) - S(H)| \\
&= \big|\mathrm{Cov}(Y_1, Y_2) - \mathrm{Cov}(Z_1, Z_2)\big| \\
&= \big|\mathrm{Cov}(Y_1, Y_2 - Z_2) - \mathrm{Cov}(Z_1 - Y_1, Z_2)\big| \\
&\leq \mathrm{Var}(Y_1)^{1/2}\mathrm{Var}(Y_2 - Z_2)^{1/2} + \mathrm{Var}(Z_2)^{1/2}\mathrm{Var}(Z_1 - Y_1)^{1/2}
\end{aligned}$$

where the last line is a consequence of the Cauchy-Schwartz inequality. We have the upper bounds

$$\mathrm{Var}(Y_1)^{1/2} \leq M_2(G), \quad \mathrm{Var}(Z_2)^{1/2} \leq M_2(H)$$

and, choosing an optimal coupling (Y, Z) between G and H,

$$\mathrm{Var}(Z_1 - Y_1)^{1/2} \leq \|Y - Z\|_{L^2} = \mathcal{W}_2(G, H), \quad \mathrm{Var}(Y_2 - Z_2)^{1/2} \leq \mathcal{W}_2(G, H).$$

Altogether, we obtain,

$$|S(G) - S(H)| \leq \big(M_2(G) + M_2(H)\big)\mathcal{W}_2(G, H).$$

This proves that S is locally Lipschitz and hence continuous with respect to the distance $\mathcal{W}_2$. Taking $H = \delta_0$, we obtain

$$|S(G)| \leq M_2(G)^2$$

and the bound (A.14) holds with $q = 2$. Thus Proposition A.2 implies that the plug-in estimator

$$S(\hat{F}_{n,x}) = \sum_{i=1}^n W_{ni}(x) Y_{1i} Y_{2i} - \sum_{i=1}^n W_{ni}(x) Y_{1i} \sum_{i=1}^n W_{ni}(x) Y_{2i}$$

is consistent in absolute mean for the conditional covariance

$$S(F_x) = \mathbb{E}(Y_1 Y_2 \mid X = x) - \mathbb{E}(Y_1 \mid X = x)\mathbb{E}(Y_2 \mid X = x),$$

i.e. $\mathbb{E}[|S(\hat{F}_{n,X}) - S(F_X)|] \longrightarrow 0$ as $n \to +\infty$.

A.4 Proofs

A.4.1 Proof of Theorem A.2

Proof of Theorem A.2 - case $d = 1$. We first consider the case when Y is uniformly bounded and takes its values in $[-M, M]$ for some $M > 0$. Then, it holds

$$F_x(z) = \begin{cases} 0 & \text{if } z < -M \\ 1 & \text{if } z \geq M \end{cases} \quad \text{and} \quad \hat{F}_{n,x}(z) = \begin{cases} 0 & \text{if } z < -M \\ 1 & \text{if } z \geq M \end{cases}.$$

and the generalized inverse functions (quantile functions) are bounded in absolute value by M. As a consequence,

$$\begin{aligned}\mathbb{E}\left[\mathcal{W}_p^p(\hat{F}_{n,X}, F_X)\right] &= \mathbb{E}\left[\int_0^1 |\hat{F}_{n,X}^{-1}(u) - F_X^{-1}(u)|^p \mathrm{d}z\right] \\ &\leq (2M)^{p-1}\mathbb{E}\left[\int_0^1 |\hat{F}_{n,X}^{-1}(u) - F_X^{-1}(u)|\mathrm{d}u\right] \\ &= (2M)^{p-1}\int_{-M}^{M} \mathbb{E}\left[|\hat{F}_{n,X}(z) - F_X(z)|\right] \mathrm{d}z. \end{aligned} \tag{A.15}$$

In these lines, we have used Equations (A.9) and (A.10) together with Fubini's theorem.

Consider the regression model $(X, \mathbb{1}_{\{Y\leq z\}}) \in \mathbb{R}^d \times \mathbb{R}$ where $z \in [-M, M]$ is fixed. The corresponding regression function is

$$x \mapsto \mathbb{E}[\mathbb{1}_{\{Y\leq z\}}|X = x] = F_x(z)$$

and the local weight estimator associated with the sample $(X_i, \mathbb{1}_{\{Y_i\leq z\}})$, $1 \leq i \leq n$ is

$$x \mapsto \sum_{i=1}^{n} W_{ni}(x)\mathbb{1}_{\{Y_i\leq z\}} = \hat{F}_{n,x}(z).$$

An application of Stone's theorem with $p = 1$ yields

$$\mathbb{E}\left[|\hat{F}_{n,X}(z) - F_X(z)|\right] \longrightarrow 0, \quad \text{as } n \to +\infty,$$

whence we deduce, by the dominated convergence theorem,

$$\int_{-M}^{M} \mathbb{E}\left[|\hat{F}_{n,X}(z) - F_X(z)|\right] \mathrm{d}z \longrightarrow 0.$$

The upper bound (A.15) finally implies

$$\mathbb{E}\left[\mathcal{W}_p^p(\hat{F}_{n,X}, F_X)\right] \longrightarrow 0.$$

We next consider the general case when Y is not necessarily bounded. For $M > 0$, we define the truncation Y^M of Y by

$$Y^M = \begin{cases} -M & \text{if } Y < -M \\ Y & \text{if } -M \leq Y < M \\ M & \text{if } Y \geq M \end{cases}.$$

We define similarly $Y_1^M, \ldots, Y_n^M$ the truncations of $Y_1, \ldots, Y_n$ respectively. The conditional distribution associated with Y^M is

$$F_x^M(z) = \mathbb{P}(Y^M \leq z|X = x) = \begin{cases} 0 & \text{if } z < -M \\ F_x(z) & \text{if } -M \leq Y < M \\ 1 & \text{if } z \geq M \end{cases}.$$

The local weight estimation built on the truncated sample is

$$\hat{F}_{n,x}^M(z) = \sum_{i=1}^{n} W_{ni}(x)\mathbb{1}_{\{Y_i^M\leq z\}}.$$

By the triangle inequality,

$$\mathcal{W}_p(\hat{F}_{n,x}, F_x) \leq \mathcal{W}_p(\hat{F}_{n,x}, \hat{F}^M_{n,x}) + \mathcal{W}_p(\hat{F}^M_{n,x}, F^M_x) + \mathcal{W}_p(F^M_x, F_x),$$

whence we deduce

$$\begin{aligned}&\mathbb{E}[\mathcal{W}^p_p(\hat{F}_{n,x}, F_x)]\\ &\leq 3^{p-1}\left(\mathbb{E}[\mathcal{W}^p_p(\hat{F}_{n,X}, \hat{F}^M_{n,X})] + \mathbb{E}[\mathcal{W}_p(\hat{F}^M_{n,X}, F^M_X)] + \mathbb{E}[\mathcal{W}^p_p(F^M_X, F_X)]\right).\end{aligned}$$

By the preceding result in the bounded case, for any fixed M, the second term converges to 0 as $n \to +\infty$. We next focus on the first and third terms.

For fixed $X = x$, there is a natural coupling between the distribution $\hat{F}_{n,x}$ and $\hat{F}^M_{n,x}$ given by (Z_1, Z_2) such that

$$(Z_1, Z_2) = (Y_i, Y_i^M) \quad \text{with probability } W_{ni}(x).$$

Clearly $Z_1 \sim \hat{F}_{n,x}$ and $Z_2 \sim \hat{F}^M_{n,x}$ and this coupling provides the upper bound

$$\mathcal{W}^p_p(\hat{F}_{n,x}, \hat{F}^M_{n,x}) \leq \|Z_1 - Z_2\|^p_{\mathrm{L}^p} = \sum_{i=1}^n W_{ni}(x)|Y_i - Y_i^M|^p. \tag{A.16}$$

Let us introduce the function $g_M(x)$ defined by

$$g_M(x) = \mathbb{E}\left[|Y - Y^M|^p \mid X = x\right].$$

Using the fact that, conditionally on $X_1, \ldots, X_n$, the random variables $Y_1, \ldots, Y_n$ are independent with distribution $F_{X_1}, \ldots, F_{X_n}$, we deduce

$$\mathbb{E}\left[\mathcal{W}^p_p(\hat{F}_{n,x}, \hat{F}^M_{n,x})\right] \leq \mathbb{E}\left[\sum_{i=1}^n W_{ni}(x) g_M(X_i)\right].$$

The condition i) on the weights in Stone's Theorem then implies

$$\mathbb{E}\left[\sum_{i=1}^n W_{ni}(X) g_M(X_i)\right] \leq C\mathbb{E}[g_M(X)].$$

Because $|Y - Y^M|^p$ converges almost surely to 0 as $M \to +\infty$ and is bounded by $2^p|Y|^p$ which is integrable, Lebesgue's convergence theorem implies

$$\mathbb{E}[g_M(X)] = \mathbb{E}\left[|Y - Y^M|^p\right] \longrightarrow 0 \quad \text{as } M \to +\infty.$$

We deduce that the first term satisfies

$$\mathbb{E}\left[\mathcal{W}^p_p(\hat{F}_{n,X}, \hat{F}^M_{n,X})\right] \leq C\mathbb{E}[g_M(X)] \longrightarrow 0, \quad \text{as } M \to +\infty$$

where the convergence is uniform in n.

We now consider the third term. Since Y^M is obtained from Y by truncation, the distribution functions and quantile functions of Y and Y^M are related by

$$F^M_x(z) = \begin{cases} 0 & \text{if } z < -M \\ F_x(z) & \text{if } -M \leq z < M \\ 1 & \text{if } z \geq M \end{cases}$$

and

$$(F_x^M)^{-1}(u) = \begin{cases} -M & \text{if } F_x^{-1}(u) < -M \\ (F_x)^{-1}(u) & \text{if } -M \le F_x^{-1}(u) < M \\ M & \text{if } F_x^{-1}(u) \ge M \end{cases}.$$

As a consequence

$$\begin{aligned} \mathcal{W}_p^p(F_x^M, F_x) &= \int_0^1 |(F_x^M)^{-1}(u) - F_x^{-1}(u)|^p \mathrm{d}u \\ &= \mathbb{E}\left[|Y^M - Y|^p \mid X = x\right] = g_M(x). \end{aligned}$$

We deduce

$$\mathbb{E}\left[\mathcal{W}_p^p(F_X^M, F_X)\right] = \mathbb{E}[g_M(X)] \longrightarrow 0, \quad \text{as } M \to +\infty$$

where the convergence is uniform in n.

We finally combine the three terms. The sum can be made smaller than any $\varepsilon > 0$ by first choosing M large enough so that the first and third terms are smaller than $\varepsilon/3$ and then choosing n large enough so that the second term is smaller than $\varepsilon/3$. This proves Equation (A.12) and concludes the proof. □

In order to extend the proof from $d = 1$ to $d \ge 2$, we need the notion of *sliced Wasserstein distance*, see Bayraktar and Guo (2021) for instance. Let $\mathbb{S}^{d-1} = \{u \in \mathbb{R}^d : \|u\| = 1\}$ be the unit sphere in $\mathbb{R}^d$ and, for $u \in \mathbb{R}^d$, let $u_* : \mathbb{R}^d \to \mathbb{R}$ be the linear form defined by $u_*(x) = u \cdot x$. The projection in direction u of a measure μ on $\mathbb{R}^d$ is defined as the pushforward $\mu \circ u_*^{-1}$ which is a measure on $\mathbb{R}$. The inequality $|u \cdot x| \le \|x\|$ implies that $\mu \circ u_*^{-1} \in \mathcal{W}_p(\mathbb{R})$ for all $\mu \in \mathcal{W}_p(\mathbb{R}^d)$ and $u \in \mathbb{S}^{d-1}$. The sliced and max-sliced Wasserstein distances between $\mu, \nu \in \mathcal{W}_p(\mathbb{R}^d)$ are then defined respectively by

$$S\mathcal{W}_p(\mu, \nu) = \left(\int_{\mathbb{S}^{d-1}} \mathcal{W}_p^p(\mu \circ u_*^{-1}, \nu \circ u_*^{-1})\, \sigma(\mathrm{d}u)\right)^{1/p},$$

where σ denotes the uniform measure on $\mathbb{S}^{d-1}$ and

$$\overline{SW}_p(\mu, \nu) = \max_{u \in \mathbb{S}^{d-1}} \mathcal{W}_p(\mu \circ u_*^{-1}, \nu \circ u_*^{-1}).$$

In plain words, the sliced and max-sliced Wasserstein distance are respectively the average and the maximum over all the 1-dimensional Wasserstein distances between the projections of μ and ν. The following result is crucial in our proof.

Theorem A.4 (Bayraktar and Guo 2021)**.** *For all $p \ge 1$, $S\mathcal{W}_p$ and $\overline{SW}_p$ are distances on $\mathcal{W}_p(\mathbb{R}^d)$ which are equivalent to $\mathcal{W}_p$, i.e. for all sequence $\mu, \mu_1, \mu_2, \ldots \in \mathcal{W}_p(\mathbb{R}^d)$*

$$S\mathcal{W}_p(\mu_n, \mu) \to 0 \iff \overline{SW}_p(\mu_n, \mu) \to 0 \iff \mathcal{W}_p(\mu_n, \mu) \to 0.$$

Proof of Theorem A.2 - case $d \ge 2$. For the sake of clarity, we divide the proof into three steps:

1) we prove that the result holds in max-sliced Wasserstein distance, i.e. $\mathbb{E}[\overline{SW}_p^p(\hat{F}_{n,X}, F_X)] \to 0$;

2) we deduce that $\mathcal{W}_p(\hat{F}_{n,X}, F_X) \to 0$ in probability;

3) we show that the sequence $\mathcal{W}_p^p(\hat{F}_{n,X}, F_X)$ is uniformly integrable.

Points 2) and 3) together imply $\mathbb{E}[\mathcal{W}_p^p(\hat{F}_{n,X}, F_X)] \to 0$ as required.

Step 1). For all $u \in \mathbb{S}^{d-1}$, the projection $\hat{F}_{n,X} \circ u_*^{-1}$ is the weighted empirical distribution

$$\hat{F}_{n,X} \circ u_*^{-1} = \sum_{i=1}^n W_{ni}(X)\delta_{Y_i \cdot u}.$$

An application of Theorem A.2 to the 1-dimensional sample $(Y_i \cdot u)_{i\geq 1}$ yields

$$\mathbb{E}[\mathcal{W}_p^p(\hat{F}_{n,X} \circ u_*^{-1}, F_X \circ u_*^{-1})] \longrightarrow 0. \tag{A.17}$$

Note indeed that $\mathbb{E}[|Y|^p] < \infty$ implies $\mathbb{E}[|Y \cdot u|^p] < \infty$ and that the conditional laws of $Y \cdot u$ are the pushforward of those of Y, i.e. $\mathcal{L}(Y \cdot u \mid X) = F_X \circ u_*^{-1}$.

We next consider the max-sliced Wasserstein distance. Regularity in the direction $u \in \mathbb{S}^{d-1}$ will be useful and we recall that the Wasserstein distance between projections depends on the direction in a Lipschitz way. More precisely, according to Bayraktar and Guo (2021, Proposition 2.2),

$$|\mathcal{W}_p(\mu \circ u_*^{-1}, \nu \circ u_*^{-1}) - \mathcal{W}_p(\mu \circ v_*^{-1}, \nu \circ v_*^{-1})| \leq (M_p(\mu) + M_p(\nu))\|u - v\|,$$

for all $\mu, \nu \in \mathcal{W}_p(\mathbb{R}^d)$ and $u, v \in \mathbb{S}^{d-1}$ (recall Equation (A.7) for the definition of $M_p(\mu)$, $M_p(\nu)$).

The sphere $\mathbb{S}^{d-1}$ being compact, for all $\varepsilon > 0$, one can find $K \geq 1$ and $u_1, \dots, u_K \in \mathbb{S}^{d-1}$ such that the balls $B(u_i, \varepsilon)$ with centers u_i and radius ε cover the sphere. Then, due to the Lipschitz property, the max-sliced Wasserstein distance is controlled by

$$\begin{aligned}
&\overline{SW}_p(\hat{F}_{n,X}, F_X) \\
&= \max_{u \in \mathbb{S}^{d-1}} \mathcal{W}_p^p(\hat{F}_{n,X} \circ u_*^{-1}, F_X \circ u^{-1}) \\
&\leq \max_{1\leq k\leq K} \mathcal{W}_p(\hat{F}_{n,X} \circ u_{k*}^{-1}, F_X \circ u_{k*}^{-1}) + \varepsilon(M_p(\hat{F}_{n,X}) + M_p(F_X)).
\end{aligned}$$

Elevating to the p-th power and taking the expectation, we deduce

$$\begin{aligned}
&\mathbb{E}\big[\overline{SW}_p^p(\hat{F}_{n,X}, F_X)\big] \\
&\leq 3^{p-1}\mathbb{E}\big[\max_{1\leq k\leq K} \mathcal{W}_p^p(\hat{F}_{n,X} \circ u_{k*}^{-1}, F_X \circ u_{k*}^{-1})\big] + 3^{p-1}\varepsilon^p(\mathbb{E}\big[M_p^p(\hat{F}_{n,X})\big] + \mathbb{E}\big[M_p^p(F_X)\big]).
\end{aligned}$$

The first term converges to 0 thanks to Eq. (A.17), i.e.

$$\mathbb{E}[\max_{1\leq i\leq K} \mathcal{W}_p^p(\hat{F}_{n,X} \circ u_{i*}^{-1}, F_X \circ u_{i*}^{-1})] \longrightarrow 0.$$

The second term is controlled by a constant times ε^p since

$$\mathbb{E}[M_p^p(\hat{F}_{n,X})] = \mathbb{E}\big[\sum_{i=1}^n W_{ni}(X)\|Y_i\|^p\big] \leq C\mathbb{E}[\|Y\|^p]$$

(by property i) of the weights) and

$$\mathbb{E}[M_p^p(F_X)] = \mathbb{E}\big[\mathbb{E}[\|Y\|^p \mid X]\big] = \mathbb{E}[\|Y\|^p]$$

(by the tower property of conditional expectation). Letting $\varepsilon \to 0$, the second term can be made arbitrarily small. We deduce $\mathbb{E}[\overline{SW}_p^p(\hat{F}_{n,X}, F_X)] \to 0$.

Step 2). As a consequence of step 1), $\overline{SW}_p(\hat{F}_{n,X}, F_X) \to 0$ in probability, or equivalently $\hat{F}_{n,X} \to F_X$ in probability in the metric space $(\mathcal{W}_p(\mathbb{R}^d), \overline{SW}_p)$. Theorem A.4 implies that the identity mapping is continuous from $(\mathcal{W}_p(\mathbb{R}^d), \overline{SW}_p)$ into $(\mathcal{W}_p(\mathbb{R}^d), \mathcal{W}_p)$. The continuous

mapping theorem implies that $\hat{F}_{n,X} \to F_X$ in probability in the metric space $(\mathcal{W}_p(\mathbb{R}^d), \mathcal{W}_p)$. Equivalently, $\mathcal{W}_p(\hat{F}_{n,X}, F_X) \to 0$ in probability.

Step 3). By the triangle inequality,

$$\mathcal{W}_p(\hat{F}_{n,X}, F_X) \leq \mathcal{W}_p(\hat{F}_{n,X}, \delta_0) + \mathcal{W}_p(\delta_0, F_X)$$

with δ_0 the Dirac mass at 0. Furthermore, for any $\mu \in \mathcal{W}_p(\mathbb{R}^d)$,

$$\mathcal{W}_p(\mu, \delta_0) = \left(\int_{\mathbb{R}^d} \|x\|^p \, \mu(\mathrm{d}x) \right)^{1/p} = M_p(\mu).$$

We deduce

$$\mathcal{W}_p^p(\hat{F}_{n,X}, F_X) \leq 2^{p-1} M_p^p(\hat{F}_{n,X}) + 2^{p-1} M_p^p(F_X).$$

In order to prove the uniform integrability of the left-hand side, it is enough to prove that

$$M_p^p(F_X) \text{ is integrable and } M_p^p(\hat{F}_{n,X}),\ n \geq 1, \text{ is uniformly integrable.} \tag{A.18}$$

We have

$$M_p^p(F_X) = \mathbb{E}[\|Y\|^p \mid X]$$

which is integrable because $\mathbb{E}[\|Y\|^p] < \infty$. Furthermore,

$$M_p^p(\hat{F}_{n,X}) = \sum_{i=1}^n W_{ni}(X) \|Y_i\|^p$$

and Stone's Theorem ensures that

$$\sum_{i=1}^n W_{ni}(X) \|Y_i\|^p \longrightarrow \mathbb{E}[\|Y\|^p \mid X] \quad \text{in } L^1.$$

Since the sequence $M_p^p(\hat{F}_{n,X})$ converges in L^1, it is uniformly integrable and the claim follows. □

A.4.2 Proof of Proposition A.1, Corollaries A.1-A.2 and Theorem A.3

Proof of Proposition A.1. The proof of the upper bound relies on a coupling argument. Without loss of generality, we can assume that the Y_i's are generated from uniform random variables U_i's by the inversion method – i.e. we assume that U_i, $1 \leq i \leq n$, are independent identically distributed random variables with uniform distribution on $(0, 1)$ that are furthermore independent of the covariates X_i, $1 \leq i \leq n$ and we set $Y_i = F_{X_i}^{-1}(U_i)$. Then the sample (X_i, Y_i) is i.i.d. with distribution P. In order to compare $\hat{F}_{n,x}$ and F_x, we introduce the random variables $\tilde{Y}_i = F_x^{-1}(U_i)$ and we define

$$\tilde{F}_{n,x}(z) = \sum_{i=1}^n W_{ni}(x) \mathbb{1}_{\{\tilde{Y}_i \leq z\}}.$$

By the triangle inequality,

$$\mathcal{W}_1(\hat{F}_{n,x}, F_x) \leq \mathcal{W}_1(\hat{F}_{n,x}, \tilde{F}_{n,x}) + \mathcal{W}_1(\tilde{F}_{n,x}, F_x).$$

In the right-hand side, the first term is interpreted as an *approximation error* comparing the weighted sample $(Y_i, W_{ni}(x))$ to $(\tilde{Y}_i, W_{ni}(x))$ where the $\tilde{Y}_i$ have the target distribution F_x. The second term is an *estimation error* where we use the weighted sample $(\tilde{Y}_i, W_{ni}(x))$ with the correct distribution to estimate F_x.

We first consider the approximation error. A similar argument as for the proof of Equation (A.16) implies

$$\mathcal{W}_1(\hat{F}_{n,x}, \tilde{F}_{n,x}) \leq \sum_{i=1}^{n} W_{ni}(x)|Y_i - \tilde{Y}_i|.$$

Introducing the uniform random variables U_i's, we get

$$\begin{aligned}
\mathbb{E}[\mathcal{W}_1(\hat{F}_{n,x}, \tilde{F}_{n,x})] &\leq \mathbb{E}\Big[\sum_{i=1}^{n} W_{ni}(x)|F_{X_i}^{-1}(U_i) - F_x^{-1}(U_i)|\Big] \\
&= \mathbb{E}\Big[\sum_{i=1}^{n} W_{ni}(x)\int_0^1 |F_{X_i}^{-1}(u) - F_x^{-1}(u)|\,\mathrm{d}u\Big] \quad \text{by independence} \\
&= \mathbb{E}\Big[\sum_{i=1}^{n} W_{ni}(x)\mathcal{W}_1(F_{X_i}, F_x)\Big],
\end{aligned}$$

where the equality relies on Equation (A.9). Note that this control of the approximation error is very general and could be extended to the Wasserstein distance of order $p > 1$.

We next consider the estimation error and our approach works for $p = 1$ only. By Equation (A.10),

$$\mathbb{E}[\mathcal{W}_1(\tilde{F}_{n,x}, F_x)] = \mathbb{E}\Big[\int_{\mathbb{R}} \Big|\sum_{i=1}^{n} W_{ni}(x)\big(\mathbb{1}_{\{\tilde{Y}_i \leq z\}} - F_x(z)\big)\Big|\mathrm{d}z\Big].$$

We apply Fubini's theorem and use the upper bound

$$\begin{aligned}
&\mathbb{E}\Big[\Big|\sum_{i=1}^{n} W_{ni}(x)\big(\mathbb{1}_{\{\tilde{Y}_i \leq z\}} - F_x(z)\big)\Big|\Big] \\
&\leq \mathbb{E}\Big[\Big|\sum_{i=1}^{n} W_{ni}(x)\big(\mathbb{1}_{\{\tilde{Y}_i \leq z\}} - F_x(z)\big)\Big|^2\Big]^{1/2} \\
&= \mathbb{E}\Big[\sum_{i=1}^{n} W_{ni}^2(x)\Big]^{1/2}\sqrt{F_x(z)(1 - F_x(z))},
\end{aligned}$$

where the last equality is obtained by integrating first with respect to $\tilde{Y}_1, \ldots, \tilde{Y}_n$ and recognizing a variance term and then with respect to $X_1, \ldots, X_n$. We deduce

$$\mathbb{E}[\mathcal{W}_1(\tilde{F}_{n,x}, F_x)] \leq \mathbb{E}\Big[\sum_{i=1}^{n} W_{ni}^2(x)\Big]^{1/2}\int_{\mathbb{R}} \sqrt{F_x(z)(1 - F_x(z))}\mathrm{d}z.$$

Collecting the two terms yields Proposition A.1. □

Proof of Corollary A.1. For the kernel algorithm with uniform kernel and weights (A.5), we denote by

$$N_n(X) = \sum_{i=1}^{n} \mathbb{1}_{\{X_i \in B(X, h_n)\}}$$

the number of points in the ball $B(X, h_n)$ with center X and radius h_n. If $N_n \geq 1$, only the points in $B(X, h_n)$ have a nonzero weight which is equal to $1/N_n$. If $N_n = 0$, then by convention all the weights are equal to $1/n$. Thus we deduce

$$\mathbb{E}\Big[\sum_{i=1}^{n} W_{ni}^2(X)\Big] = \mathbb{E}\Big[\frac{1}{N_n(X)}\mathbb{1}_{\{N_n(X) \geq 1\}}\Big] + \frac{1}{n}\mathbb{P}(N_n(X) = 0)$$

and

$$\mathbb{E}\Big[\sum_{i=1}^{n} W_{ni}(X)\|X_i - X\|^H\Big] \leq h_n^H \mathbb{P}(N_n(X) \geq 1) + k^{H/2}\mathbb{P}(N_n(X) = 0)$$

because the distance to X for the points with nonzero weight can be bounded from above by h_n if $N_n(X) \geq 1$ and by $\sqrt{k}$ otherwise (note that $\sqrt{k}$ is the diameter of $[0,1]^k$).

Next, we use the fact that, conditionally on $X = x$, $N_n(x)$ has a binomial distribution with parameters n and $p_n(x) = \mathbb{P}(X_1 \in B(x, h_n))$. This implies

$$\mathbb{E}\Big[\frac{1}{N_n(X)}\mathbb{1}_{\{N_n(X)\geq 1\}}\Big] \leq \mathbb{E}\Big[\frac{2}{np_n(X)}\Big] \leq \frac{2c_k}{nh_n^k}$$

where the first inequality follows from Györfi et al. (2002, Lemma 4.1) and the second one from Györfi et al. (2002, Equation 5.1) where the constant $c_k = k^{k/2}$ can be taken. Similarly,

$$\begin{aligned}
\mathbb{P}(N_n(X) = 0) = \mathbb{E}[(1 - p_n(X))^n] &\leq \mathbb{E}[e^{-np_n(X)}] \\
&\leq \big(\max_{u>0} ue^{-u}\big) \times \mathbb{E}\Big[\frac{1}{np_n(X)}\Big] \\
&\leq \frac{c_k}{nh_n^k}.
\end{aligned}$$

In view of these different estimates, Equation (A.13) entails

$$\begin{aligned}
\mathbb{E}\big[\mathcal{W}_1(\hat{F}_{n,X}, F_X)\big] &\leq L\Big(h_n^H + k^{H/2}\frac{c_k}{nh_n^k}\Big) + M\left(\frac{(2+1/n)c_k}{nh_n^k}\right)^{1/2} \\
&\leq Lh_n^H + M\sqrt{(2+1/n)c_k}(nh_n^k)^{-1/2} + Lk^{H/2}c_k(nh_n^k)^{-1}.
\end{aligned}$$

□

Proof of Corollary A.2. For the nearest neighbor weights (A.6), there are exactly κ_n non-vanishing weights with value $1/\kappa_n$ whence

$$\sum_{i=1}^{n} W_{ni}^2(X) = \frac{1}{\kappa_n}.$$

Furthermore, the κ_n nearest neighbors of X satisfy

$$\|X_{i:n}(X) - X\| \leq \|X_{\kappa_n:n}(X) - X\|, \quad i = 1, \ldots, \kappa_n.$$

In view of this, Equation (A.13) entails

$$\begin{aligned}
\mathbb{E}\big[\mathcal{W}_1(\hat{F}_{n,X}, F_X)\big] &\leq L\mathbb{E}\big[\|X_{\kappa_n:n}(X) - X\|^H\big] + M\kappa_n^{-1/2} \\
&\leq L\mathbb{E}\big[\|X_{\kappa_n:n}(X) - X\|^2\big]^{H/2} + M\kappa_n^{-1/2}
\end{aligned}$$

where the last line relies on Jensen's inequality. We conclude thanks to Biau and Devroye (2015, Theorem 2.4) stating that

$$\mathbb{E}\left[\|X_{\kappa_n:n}(X) - X\|^2\right] \leq \begin{cases} 8(\kappa_n/n) & \text{if } k = 1, \\ \tilde{c}_k(\kappa_n/n)^{2/k} & \text{if } k \geq 2. \end{cases}$$

□

Proof of Theorem A.3 (lower bound). The proof of a lower bound for the minimax risk in Wasserstein distance is adapted from the proof of Proposition 3 in Pic et al. (2023, Appendix C) and we give only the main lines.

Consider the subclass of $\mathcal{D}(H, L, M)$ where Y is a binary variable with possible values 0 and B. Note that condition c) of Definition A.1 is automatically satisfied if $B \leq 4M$. The conditional distribution of Y given $X = x$ is characterized by

$$p(x) = \mathbb{P}(Y = B \mid X = x)$$

and the Wasserstein distance by

$$\mathcal{W}_1(F_x, F_{x'}) = B|p(x) - p(x')|,$$

so that property b) of Definition A.1 is equivalent to

$$B|p(x) - p(x')| \leq L\|x - x'\|^H. \tag{A.19}$$

Similarly as in Pic et al. (2023, Lemma 1), one can show that a general prediction with values in $\mathbb{R}$ can always be improved (in terms of Wasserstein error) into a binary prediction with values in $\{0, B\}$. Indeed, for a given prediction $\hat{F}_{n,x}$, the binary prediction

$$\tilde{F}_{n,x} = (1 - \tilde{p}_n(x))\delta_0 + \tilde{p}_n(x)\delta_B$$

with

$$\tilde{p}_n(x) = \frac{1}{B}\int_0^B \big(1 - \hat{F}_{n,x}(z)\big)\mathrm{d}z$$

always satisfies

$$\mathbb{E}[\mathcal{W}_1(\tilde{F}_{n,X}, F_X)] \leq \mathbb{E}[\mathcal{W}_1(\hat{F}_{n,X}, F_X)].$$

This simple remark implies that, when considering the minimax risk on the restriction of the class $\mathcal{D}(H, L, M)$ to binary distributions, we can focus on binary predictions. But for binary predictions,

$$\mathbb{E}[\mathcal{W}_1(\tilde{F}_{n,X}, F_X)] = B|\tilde{p}_n(X) - p(X)|,$$

showing that the minimax rate of convergence for distributional regression in Wasserstein distance is equal to the minimax rate of convergence for estimating the regression function $\mathbb{E}[Y|X = x] = Bp(x)$ in absolute error under the regularity assumption (A.19). According to Stone (1980, 1982), a lower bound for the minimax risk in L^1-norm is $n^{-H/(2H+k)}$ (in the first paper, we consider the Bernoulli regression model referred to as Model 1 Example 5 and the L^q distance with $q = 1$). □

Proof of Theorem A.3 (upper bound). For the kernel method, Corollary A.1 states that the expected Wasserstein error is upper bounded by

$$Lh_n^H + M\sqrt{(2 + 1/n)c_k}(nh_n^k)^{-1/2} + Lk^{H/2}c_k(nh_n^k)^{-1}.$$

Minimizing the sum of the first two terms in the right-hand side with respect to h_n leads to $h_n \propto n^{-1/(2H+k)}$ and implies that the right-hand side is of order $n^{-H/(2H+k)}$ (the last term is negligible). This matches the minimax lower rate of convergence previously stated previously and proves that the optimal minimax risk is of order $n^{-H/(2H+k)}$.

For the nearest neighbor method, minimizing the upper bound for the expected Wasserstein error from Corollary A.2 leads to

$$\kappa_n \propto \begin{cases} n^{H/(H+1)} & \text{if } k = 1 \\ n^{H/(H+k/2)} & \text{if } k \geq 2 \end{cases},$$

with a corresponding risk of order

$$\begin{cases} n^{-H/(2H+2)} & \text{if } k = 1 \\ n^{-H/(2H+k)} & \text{if } k \geq 2 \end{cases},$$

whence the nearest neighbor method reaches the optimal rate when $k \geq 2$. □

A.4.3 Proof of Proposition A.2

Proof of Proposition A.2. The first point follows from the fact that composition by a continuous application respects convergence in probability. Indeed, as the estimator $\hat{F}_{n,X}$ converges to F_X in probability for the Wasserstein distance $\mathcal{W}_p$, $S(\hat{F}_{n,X})$ converges to $S(F_X)$ in probability.

In order to prove the consistency in $\mathrm{L}^{p/q}$, it is enough to prove furthermore the uniform integrability of $|S(\hat{F}_{n,X}) - S(F_X)|^{p/q}$, $n \geq 1$. With the convexity inequality of power functions as $p/q \geq 1$, Equation (A.14) entails

$$\begin{aligned} |S(\hat{F}_{n,X}) - S(F_X)|^{p/q} &\leq 2^{p/q-1}\big(|S(\hat{F}_{n,X})|^{p/q} + |S(F_X)|^{p/q}\big) \\ &\leq 2^{p/q-1}\Big((aM_p^q(\hat{F}_{n,X}) + b)^{p/q} + (aM_p^q(F_X) + b)^{p/q}\Big) \\ &\leq 2^{2(p/q-1)}\Big(a^{p/q}M_p^p(\hat{F}_{n,X}) + a^{p/q}M_p^p(F_X) + 2b^{p/q}\Big). \end{aligned}$$

This upper bound together with Equation (A.18) implies the uniform integrability of $|S(\hat{F}_{n,X}) - S(F_X)|^{p/q}$, $n \geq 1$, which concludes the proof. □

Acknowledgements

The authors acknowledge the support of the French Agence Nationale de la Recherche (ANR) under reference ANR-20-CE40-0025-01 (T-REX project). They are also grateful to Mehdi Dagdoug for suggesting the example of random forest weights (Example A.2).

Appendix B

A new methodology to predict the oncotype scores based on clinico-pathological data with similar tumor profiles

This chapter reproduces an article published in *Breast Cancer Research and Treatment*, and written by Zeina Al Masry[1], Romain Pic[2], Clément Dombry[2] and Christine Devalland[3].

Abstract

Purpose The Oncotype DX (ODX) test is a commercially available molecular test for breast cancer assay that provides prognostic and predictive breast cancer recurrence information for hormone-positive, HER2-negative patients. This study aims to propose a novel methodology to assist physicians in their decision-making.

Methods A retrospective study between 2012 and 2020 with 333 cases that underwent an ODX assay from three hospitals in the Bourgogne Franche-Comté region (France) was conducted. Clinical and pathological reports were used to collect the data. A methodology based on distributional random forests was developed to predict the ODX score classes (ODX ≤ 25 and ODX > 25) using 9 clinico-pathological characteristics. This methodology can be used to identify the patients of the training cohort that share similarities with the new patient and to predict an estimate of the distribution of the ODX score.

Results The mean age of participants is 56.9 years old. We have correctly classified 92% of patients in low risk and 40.2% of patients in high risk. The overall accuracy is 79.3%. The proportion of low-risk correct predicted value (PPV) is 82%. The percentage of high-risk correct predicted value (NPV) is approximately 62.3%. The F1-score and the Area Under the Curve (AUC) are 0.87 and 0.759, respectively.

Conclusion The proposed methodology makes it possible to predict the distribution of the ODX score for a patient. The determination of a family of known patients with a follow-up of identical scores reinforces this prediction. Using this methodology with the pathologist's expertise on the different histological and immunohistochemical characteristics has a clinical impact to help oncologists in decision-making regarding breast cancer therapy.

[1]Institut FEMTO-ST, Université Bourgogne Franche-Comté, CNRS, SUPMICROTECH-ENSMM, Besançon, France

[2]Université de Franche Comté, CNRS, LmB (UMR 6623), F-25000 Besançon, France

[3]Service d'anatomie et cytologie pathologiques, Hôpital Nord Franche-Comté, 90400 Trévenans, France

Contents

B.1 Introduction

The Oncotype DX (ODX) test is a commercially available molecular test for breast cancer assay (Genomic Health) that provides prognostic and predictive breast cancer recurrence information for hormone-positive, HER2-negative patients. The ODX test is based on 1A-level evidence and it is included in the main international clinical guidelines recommendations such as those of the American Society of Clinical Oncology (ASCO; Andre et al., 2019) or the National Comprehensive Cancer Network (NCCN) as well as in the last staging guidelines of AJCC 8th edition (Giuliano et al., 2017). The ODX test is the most widely available molecular test used in the world. This assay analyzes 21 genes by RT-qPCR (16 cancer-related genes and 5 housekeeping genes) and aims to predict the risk of recurrence at 10 years by providing a recurrence score ranging from 0 to 100 and to estimate the benefit of adjuvant chemotherapy. Several retrospective and prospective studies have validated this test and its clinical utility. Paik et al. (2004) have shown a correlation between ODX score and disease-free survival in patients with ER-positive/HER2-negative, node-negative, tamoxifen-treated breast cancer, based on the NSABP B-14 trial. As for the chemotherapy benefits, Paik et al. (2006) and Albain et al. (2010) have evaluated the test using studies related to NSABP-B20 and SWOG 8814. The prospective phase III trial TAILORx study (Sparano et al., 2018) has modified the ODX score's cutoff values (low risk <11, intermediate risk 11-25 and high risk >25) in order to avoid under-treatments of cancer. To be more precise, in the low-risk group, the risk of recurrence at 5 years is very low (<10%) with hormonal therapy, which confirms the uselessness of adding chemotherapy (Sparano et al., 2015). For the intermediate group, chemotherapy has a benefit only for women younger than 50 years old. For the high-risk group, chemotherapy is highly recommended. Nevertheless, one-third of women with hormone-receptor-positive breast cancer have a lymph node disease. Thus, the prospective trial RxPONDER trial study analyzes the capacity of the ODX test to predict the benefit of chemotherapy for women with positive lymph node disease (Kalinsky et al., 2021). RxPONDER showed that postmenopausal patients with node involvement and an ODX score between 0 and 25 did not benefit from chemotherapy, whereas premenopausal patients with node involvement with 1-3 nodes and ODX scores between 0 and 25 benefited significantly from chemotherapy.

Despite its proven value, the ODX test is not routinely used due to its high cost. For this reason, less than 20% of patients in Europe have access to the ODX test. Health-related economic studies are performed to understand for which patients the assay is the most useful (Albanell et al., 2016). From this economic point of view, many alternative tools have been developed to predict this score. These tools are based on clinico-pathological data such as Magee equations (Klein et al., 2013; Sughayer et al., 2018) and the IHC4 score (Yeo et al., 2015). Indeed, many studies have shown the correlation between the results of the latter tools

and the ODX score (see, e.g., Flanagan et al., 2008). Few works used features with machine learning techniques in order to provide an ODX-based methodology to divide the patients into categories corresponding to low or high risk of cancer (Kim et al., 2019; Orucevic et al., 2019; Baltres et al., 2020; Pawloski et al., 2021).

This paper aims to propose a novel methodology to assist physicians in their decision-making. It is based on random forests for distributional regression as presented in Meinshausen (2006) and Athey et al. (2019). This methodology creates links between a new patient and the cohort used for training based on clinico-pathological characteristics. These links can be used particularly to identify the patients of the training cohort that share similarities with the new patient and to predict an estimate of the distribution of the ODX score. This information is available to clinicians to help them better understand the probable clinical evolution of the tumor in order to optimize the treatment.

Moreover, it enables knowledge capitalization by feedback and analysis of patient history. One of the consequences of this study is to weigh the variability of the anatomo-pathological data so that this new methodology can adapt to the specificities of a cohort.

B.2 Materials and methods

B.2.1 Dataset description

The cohort is a retrospective study between 2012 and 2020 with 333 cases that underwent an ODX assay from three hospitals in Bourgogne Franche-Comté: Besançon, Belfort and Dijon. All patients have ER-positive and HER2-negative early breast cancer. Clinical and pathological reports were used to collect the data such as the age at diagnosis, the menopausal status, the treatment, the recurrence, the tumor size, the lymph node status, the histological type, the Nottingham grade, hormone receptors for estrogen (ER) expression, hormone receptors for progesterone (PR) expression, the human epidermal growth factor receptor 2 (HER2) status and the protein p53 and Ki67 proliferation index. Immunohistochemical staining was performed (Ventana Benchmark XT system®, Roche™) on the tumor block of ODX testing with UltraView Universal DAB detection with ER antibody (clone SP1; Roche/Ventana Medical Systems, Tucson, USA), PR antibody (clone 1E2; Roche/Ventana Medical Systems, Tucson, USA), HER2 antibody (clone 4B5; Roche/Ventana Medical Systems, Tucson, USA), Ki67 antibody (clone Mib-1, Dako, Glostrup, Denmark) and p53 antibody (clone DO-7, Dako, Glostrup, Denmark). The HER2 immunostaining was interpreted using the 2018 American Society of Clinical Oncology/College of American Pathologists guidelines (Wolff et al., 2018). The Ki67 proliferation index was evaluated by manual counting with a counter on at least 200 tumor cells. The protein p53 was assessed by immunohistochemistry. The positive threshold is greater than 10% of the tumor cells' nuclei. The ODX test was realized by Genomic Health (Redwood City, CA, USA) and analyzed 21 genes by RT-qPCR from paraffin-embedded blocks of tumor tissue. The ODX score was obtained from the clinical reports. The three ODX categories were the same as the ones defined in the ODX's assays using TAILORx and RxPONDER: low risk (<16), intermediate risk (16-25) and high risk (>25). The institution review board approved this study.

The cohort contains more than 50 features, from which we selected the most critical ones using feature importance in random forest and physicians' assessments. Table B.1 describes the tumor characteristics using the features selected for our study.

B.2.2 Distributional Random Forest

Random Forest (Breiman, 2001) is a powerful machine learning algorithm that can be used for prediction in various settings and has been successfully applied in the field of medicine (Chen et al., 2020; Fernandez-Lozano et al., 2021; Zare et al., 2021). Our goal here is to predict the

		Percentage of patient by category			
		< 16	$16 - 25$	> 25	Total
Population		113	138	82	333
Age	≤ 50 yr	14.41	10.51	5.71	30.63
	> 50 yr	19.52	30.92	18.92	69.37
Tumor size	< 1 cm	3.90	4.51	3.00	11.41
	1-2 cm	15.62	22.52	15.62	53.76
	> 2 cm	14.41	14.41	6.01	34.83
p53	$\leq 10\%$	18.62	23.12	12.01	53.75
	$> 10\%$	15.32	18.32	12.61	46.25
SBR grade	1	5.41	3.30	0.00	8.71
	2	21.02	24.03	10.81	55.86
	3	7.51	14.11	13.81	35.43
Mitotic grade	1	12.61	14.42	4.50	31.53
	2	17.12	20.12	12.01	49.25
	3	4.20	6.91	8.11	19.22
ER status	Negative	0.00	0.00	0.00	0.00
	Positive ($\geq 10\%$)	33.93	41.44	24.63	100
PR status	Negative	2.10	7.51	8.11	17.72
	Positive ($\geq 10\%$)	31.83	33.93	16.52	82.28
Ki67-positive cells	$< 10\%$	0.00	0.30	0.00	0.30
	$10 - 20\%$	16.22	15.92	4.80	36.94
	$> 20\%$	17.72	25.22	19.82	62.76
Lymph node status	0	15.02	15.52	13.81	45.35
	1	10.81	14.11	4.20	29.13
	2	3.61	3.30	0.90	7.81
	3	1.80	2.10	2.10	6.00
	NA	2.70	5.41	3.60	11.71

Table B.1: ODX score distribution by patient and tumor characteristics.

result of the expensive ODX test based on clinico-pathological features. We propose the use of a Distributional Random Forest that provides a predictive distribution for the ODX score based on the clinico-pathological features. We shall expose the methodology for Random Forest and Distributional Random Forest.

Standard regression links the mean of the response variable Y to a set of features X based on observations from a training sample of feature–response pairs, say (X_i, Y_i) for $i = 1, \ldots, n$. Random Forest (RF) prediction is an ensemble method that consists of the bootstrap aggregation (Breiman, 1996) of randomized classification and regression trees (CART, Breiman et al., 1984). The predictive mean can be written as the average

$$\hat{Y} = \frac{1}{B} \sum_{b=1}^{B} T_b(X), \tag{B.1}$$

where $T^1(X), \ldots, T^B(X)$ corresponds to the prediction of the different trees built on different bootstrap samples. Each single tree prediction takes the form of an average across a neighborhood of X in the tree, i.e.

$$T_b(X) = \frac{1}{|R_b(X)|} \sum_{X_i \in R_b(X)} Y_i,$$

with $R_b(X)$ being the region of the feature space that contains X in the tree T_b and $|R_b(X)|$ the numbers of observations that fall into this region. Consequently, the Random Forest prediction

(B.1) has the equivalent form

$$\hat{Y} = \sum_{i=1}^{n} w_i(X) Y_i, \tag{B.2}$$

with the Random Forest weights defined by

$$w_i(X) = \frac{1}{B} \sum_{i=1}^{B} \frac{\mathbb{1}_{\{X_i \in R(X)\}}}{|R_b(X)|}, \quad 1 \leq i \leq n, \tag{B.3}$$

and these weights are non-negative with sum 1 (probability weights).

The main idea of Distributional Random Forest (DRF) relies on Equation (B.2): the prediction $\hat{Y}$ is the sample mean of the weighted sample Y_i with weights $w_i(X)$ which can be seen as an approximation of the conditional distribution of Y given X. The cumulative distribution function $F(y|X) = \mathbb{P}(Y \leq y|X)$ is thus approximated by

$$\hat{F}(y|X) = \sum_{i=1}^{n} w_i(X) \mathbb{1}_{\{Y_i \leq y\}}. \tag{B.4}$$

This idea was first suggested by Meinshausen (2006) who proposed the construction of a quantile regression forest by approximating the conditional quantile of Y given X by the quantiles of the weighted empirical distribution (B.4).

Figure B.1 presents a synthetic representation of the DRF procedure with the different steps: subsampling of the original sample, tree construction on each subsample, computation of the neighborhood/weight at the point to predict, averaging of weights given by the different trees that finally provide the predictive distribution.

The Random Forest weights (B.3) are interesting in themselves and provide relevant information in terms of similar/influential observations. Given a new feature X, the weight $w_i(X)$ is interpreted as the proportion in which the observation Y_i contributes to the prediction of Y given X. Observations with the largest weights are interpreted as the nearest neighbors of X in terms of an implicit metric on the predictor space that is tailored for predicting the response, see Lin and Jeon (2006). The random forest weights make it possible to identify the most similar/influential individuals in the training data. Comparing X to these similar observations can help understand the relationship between X and Y.

Finally, let us mention that the weights (B.2)-(B.3) depend on the specific structure of the trees that are used for prediction. Trees are grown by recursive binary splitting, maximizing a homogeneity criterion; the goal is to partition the feature space into different regions that are as homogeneous as possible. In the standard CART algorithm, the variance is used as the homogeneity criterion, resulting in a partition adapted to the prediction of the mean. Several different splitting rules have been considered in the statistical literature that emphasized the prediction of quantiles (Athey et al., 2019, Generalized Random Forest) or on the overall distribution (Ćevid et al., 2022, Distributional Random Forest).

A Distributional Random Forest is fitted to the whole data set. The software R with the package `grf` (Generalized Random Forest) is used to compute the random forest and the associated weights. When no new test set is provided, the `grf::predict` method performs out-of-bag prediction on the training set. This means that, for each training example, all the trees that did not use this example during the training are identified (the example was 'out-of-bag'), and a prediction for the test example is then made using only these trees.

B.2.3 Applications of Distributional Random Forest

DRF is a fully non-parametric and model-free method that performs probabilistic forecast and distributional regression. For a set of features X, it provides the full predictive distribution

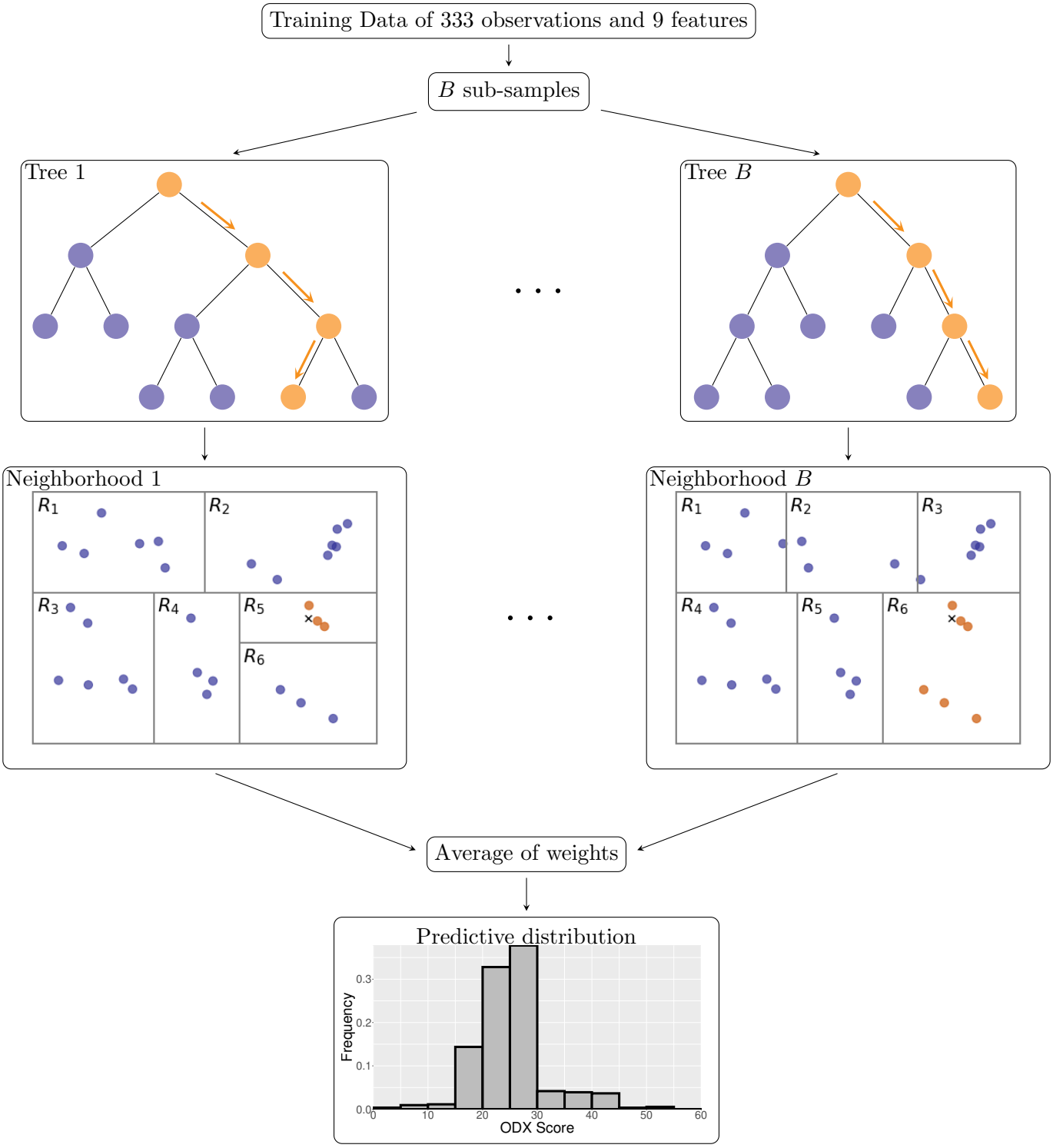
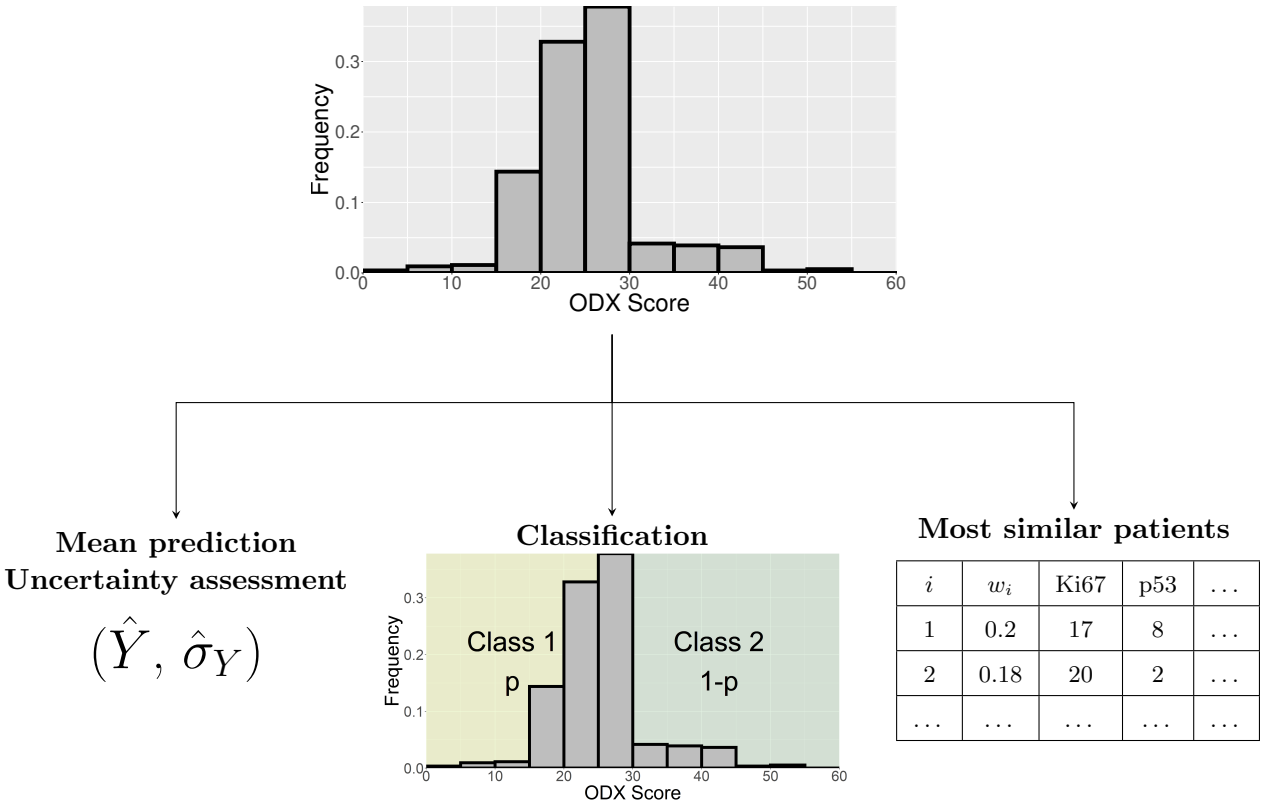


Figure B.1: Flowchart of Distributional Random Forest. Starting from the training data, a large number of subsamples are randomly chosen and binary trees are constructed on each subsample; the neighborhood/weights at the point to predict are computed in each tree and then averaged to give the forest weights; the predictive distribution corresponds to the weighted sample of the original training data with these forest weights.

of the response variable Y, that is to say, exhaustive information for its possible fluctuations knowing the features. The method is very informative and powerful (see Figure B.2) as it provides:

- (distributional regression) a predictive distribution for each new case that can be represented by a histogram;
- (mean or median prediction) a predictive mean or median when a point estimate is needed - the mean is commonly used while the median is more robust to outliers;
- (uncertainty assessment) a graphical assessment of the uncertainty with the shape of the histogram (either peaked or flat) or numerical statistics such as standard error or confidence interval for the prediction;
- (classification) the probability of classes of particular interest can be instantly computed - for the ODX score, the classes ODX ≤ 25 and ODX > 25 are considered;
- (similar/influential patients) the patients in the cohort (training set) that are the most similar to a new case can be easily identified through the random forest weights that are interpreted as a measure of proximity - this proximity is meant in the sense of an implicit distance that is learned by the model and that gives more importance to the relevant features; this information can allow the practitioner to make meaningful and informative comparisons between the new case and the patients from the cohort.



i	w_i	Ki67	p53	...
1	0.2	17	8	...
2	0.18	20	2	...
...	...	...	...	...

Figure B.2: Applications of Distributional Random Forest. Once the DRF is trained, the prediction of classes (classification) or conditional mean or median (regression) together with an uncertainty estimate is straightforward. Furthermore, the weights at the point to predict make it possible to identify the most similar neighbors in the training data, with an adaptive notion of similarity tailored for the purpose of prediction.

B.2.4 Evaluation of predictive performance

To evaluate the distributional random forest algorithm and compare it with concurrent methods, the theory of proper scoring rules (Gneiting and Raftery, 2007) is used. In probabilistic forecasting, a scoring rule compares a predictive distribution F and the outcomes y. It plays the role of a measure of error similar to the mean squared error in regression or the misclassification rate in classification. A scoring rule is strictly proper if the expected score is minimal when the predictive distribution F matches the outcome distribution. A strictly proper scoring rule can be used for the evaluation of probabilistic forecast and distributional regression (Gneiting and Katzfuss, 2014).

The most popular scoring rule is the Continuous Ranked Probability Score (CRPS; Matheson and Winkler, 1976) and is defined by

$$\mathrm{CRPS}(F, y) = \int_{\mathbb{R}} (F(z) - \mathbb{1}_{\{y \leq z\}})^2 \mathrm{d}z.$$

In a case where the predictive distribution F corresponds to a weighted sample $(y_i)_{1 \leq i \leq n}$ with weights $(w_i)_{1 \leq i \leq n}$, the CRPS is easily computed by

$$\mathrm{CRPS}(F, y) = \sum_{i=1}^{n} w_i |y_i - y| - \sum_{1 \leq i < j \leq n} w_i w_j |y_i - y_j|.$$

The first term compares the predictive distribution F and observation y (calibration) while the second term assesses the precision of the prediction (sharpness). This expression also shows that $\mathrm{CRPS}(F, y)$ is reported in the same unit as the observation y and that it generalizes the absolute error to which it is reduced if F is a deterministic forecast, that is to say a point measure.

In order to evaluate the generalization capacity of the model, that is to say, its predictive performance on a new sample, different validation methods can be used to assess the prediction error. Simple validation uses a training set to fit the model and an independent test set to compute error (CRPS). K-fold cross-validation is more involved and splits the data into K groups that successively play the role of the test set. More precisely, K different models are fitted on training sets consisting of all folds but one which is left out during training and used as a test set to compute the CRPS; this results in K different test errors which are averaged so as to obtain the K-fold cross-validation error. In the specific case of bagging including our random forest method, the out-of-bag (OOB) method can be used instead. It usually provides similar results as K-fold cross-validation but is much more numerically efficient since only one fit of the model is required. Indeed, due to resampling, a given observation does not belong to all the subsamples and one can consider the submodel aggregating all the trees that were trained without this observation; this submodel is then evaluated at the observation and the error (CRPS) is computed; averaging all these errors yield the OOB error.

B.3 Results

The DRF was applied to 333 patients to predict the ODX score using the 9 features presented in Table B.1. In order to compare with the literature, we emphasize the classification into two classes ($\mathrm{ODX} \leq 25$ and $\mathrm{ODX} > 25$).

Before presenting the results of the DRF, we shall first present the evaluation of our model. Simple graphical diagnostics can be performed by considering the regression model deduced from DRF. The results of the regression are presented in Figure B.3, where the predictive mean (Figure B.3a) and predictive median (Figure B.3b) versus the real ODX score are plotted. We can observe a rather good fit, and that an important proportion of the observations are within

their confidence intervals. In Figure B.3a, the grey ribbon has a semi-amplitude equal to the standard error and accounting for uncertainty. The grey ribbon in Figure B.3b represents the credibility interval with a level of 90%.

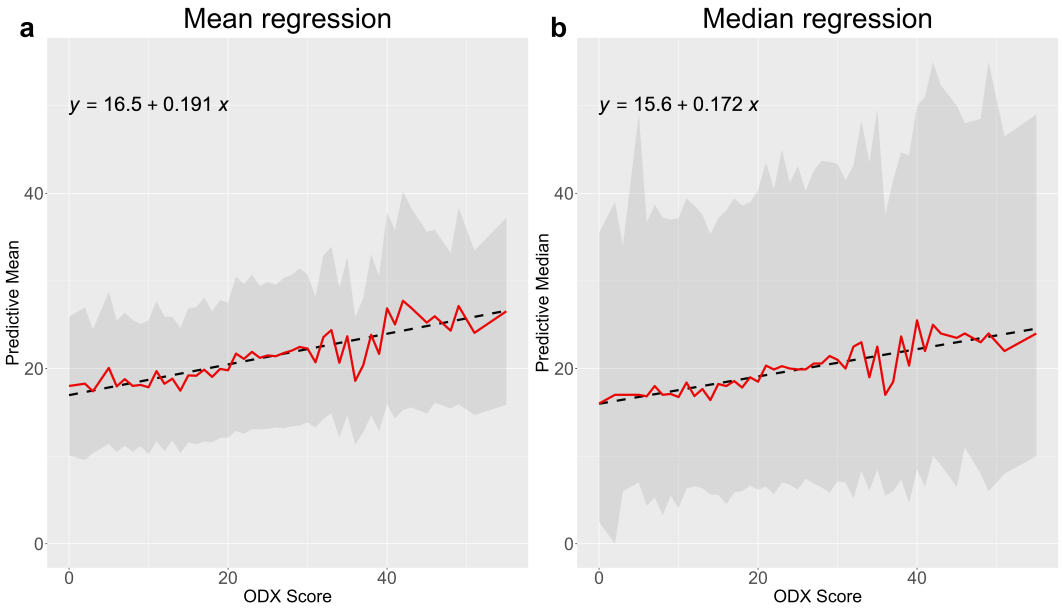


Figure B.3: Evaluation with regression diagnostics. Mean regression (left figure - a) plots the ODX observation versus their predictive mean; the grey ribbon represents the standard errors. Median regression (right figure - b) plots the ODX score versus their predictive median; the grey ribbon represents the 90%-confidence interval.

Additionally, in order to assess the ODX probabilistic forecast, we compared the OOB predictive distribution and the actual observation for the ODX, for each observation. The prediction error is measured in terms of the CRPS introduced in Section B.2.4. The different scores are represented in Figure B.4a. The smaller the CRPS, the more accurate the forecast. We can observe that most of the predictions have a small or medium CRPS, which indicates the overall good quality of prediction. A smaller number of observations have a large CRPS, indicating individuals for whom the ODX score notably differs from what we might expect in comparison with the overall population. Together with the CRPS, the figure provides the results for the binary classification task (ODX ≤ 25 or ODX > 25): classification errors are indicated with the color orange while the color blue corresponds to correctly classified observations. We can observe a good match between classification errors and a large CRPS, which confirms the ability of the CRPS to assess forecast quality. Then, for each patient, the DRF provides a predictive distribution represented by a histogram that can be compared with the actual ODX score. We also indicate the two class probabilities corresponding to the light-green/left or dark-green/right classes.

We have selected three patients respectively with a low (Figure B.4b, Patient A), medium (Figure B.4c, Patient B) and large CRPS (Figure B.4d, Patient C). The predictions associated with these patients can be considered "good", "average" and "bad", respectively. In Figure B.4b we can observe a sharp predictive distribution (peaked histogram) and an ODX score close to the peak. In Figure B.4c, the histogram is flatter, indicating more uncertainty, and the true ODX score is contained in a high probability region. In Figure B.4d, the predictive distribution has also a large dispersion and the ODX score is contained in a low probability region, which means that the match between the two is poor. We insist on the fact that a large CRPS does not necessarily mean a miss-classification of a patient as it can be seen for some of the higher CRPS values in Figure B.4a. The CRPS considers the distributional regression and is not explicitly

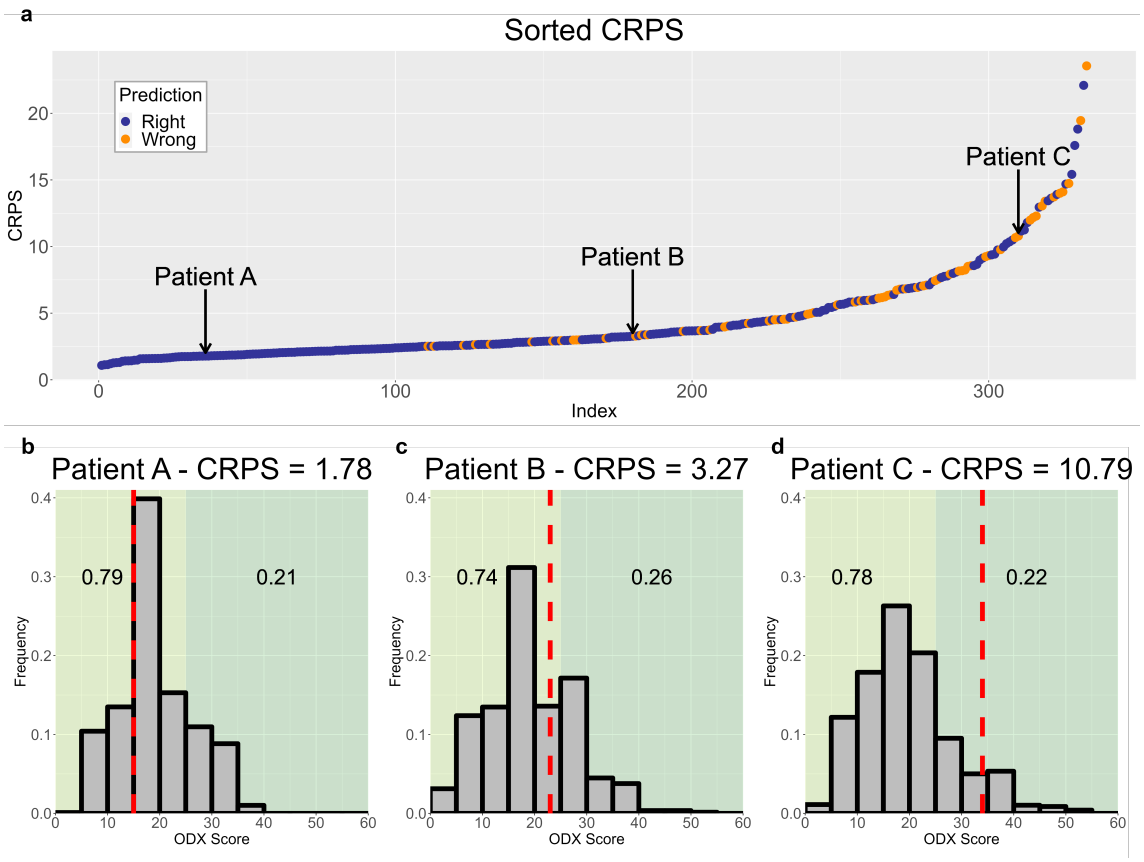


Figure B.4: Out-of-Bag evaluation of the prediction with the CRPS - a low CRPS corresponds to a precise forecast. The subfigures b, c and d correspond to three different patients chosen in different ranges of the CRPS presented in the subfigure a. In the three lower subfigures, the gray histogram corresponds to the predicted distribution of the ODX score obtained by the DRF. The red dashed line represents the true ODX score of the patient. The two classes (ODX ≤ 25 and ODX > 25) are represented as areas of different colors and the predicted probabilities of each class are given for each patient.

related to the binary classification presented here.

Due to the impact of the classification of ODX in the two classes ODX ≤ 25 and ODX > 25, we shall present the detailed evaluation of the classification model deduced from DRF (see Table B.2). This evaluation is based on the standard classification metrics such as the confusion matrix and standard metrics. The standard metrics are as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN}, \tag{B.5}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN}, \tag{B.6}$$

$$\text{Specificity} = \frac{TN}{FP + TN}, \tag{B.7}$$

$$\text{Positive Predictive Value} = \frac{TP}{TP + FP}, \tag{B.8}$$

$$\text{Negative Predictive Value} = \frac{TN}{FN + TN}, \tag{B.9}$$

		Predicted	
		ODX ≤ 25	ODX > 25
True	ODX ≤ 25	231	20
	ODX > 25	49	33

Accuracy	79.3%
Sensitivity	92.0%
Specificity	40.3%
Positive Predictive Value	82.5%
Negative Predictive Value	62.3%
F1-score	0.870
Area Under Curve	0.759

Table B.2: Evaluation with classification diagnostics. Confusion matrix (left) together with standard metrics (right).

$$\text{F1-score} = \frac{2 * \text{Positive Predictive Value*Sensitivity}}{\text{Positive Predictive Value+Sensitivity}} \tag{B.10}$$

where TP is the number of patients correctly classified as ODX ≤ 25, FP is the number of patients incorrectly classified as ODX ≤ 25, TN is the number of patients correctly classified as ODX > 25 and FN is the number of patients incorrectly classified as ODX > 25.

We have correctly classified 231 out of 251 patients (92%) as low-risk and 33 of 75 patients (40.2%) as high-risk. The overall accuracy is 79.3% and the p-value is less than 0.05. The proportion of low-risk correct predicted value (PPV) is 82%. The percentage of high-risk correct predicted value (NPV) is approximately 62.3%. The F1-score and the Area Under the Curve (AUC) are 0.87 and 0.759, respectively. The DRF will provide additional information such as the nearest neighbor patients, the distribution of the ODX score and the uncertainty prediction (see Figure B.2). We now consider the 69 miss-classified patients with low and high risks. First of all, we notice that the majority of these patients have predictions that are close to the decision border (i.e. close to ODX = 25). These patients are miss-classified because of the binary decision and additional information available with the DRF method shows either that the patient's ODX score is close to the decision border or that the neighborhood of the patient is not realistic because of limitations of the training cohort. This first part of the miss-classified patients might have a small CRPS as the CRPS accounts for the dispersion of the prediction and its bias. The second part of the miss-classified patients corresponds to extreme values of the ODX score within our cohort. The nearest patients provided by the DRF for these miss-classified patients are thus less informative as they are taken within the cohort that is not representative of these outlier patients. In order to give more quantitative results, we compared the mean absolute difference for the ODX score, Ki67 and p53 between the 69 miss-classified patients and the weighted average value of their neighborhoods. The miss-classified patients have a mean absolute difference of ODX score compared to their neighborhood of 9.84 whereas the correctly classified patients have an average absolute difference of 6.29. In terms of Ki67 and p53, the average absolute difference is 24.56% and 5.77% respectively when the average absolute difference for the correctly classified patient is 16.77% for Ki67 and 5.84% for the p53 respectively.

These classification results are then compared with state-of-the art techniques: Klein et al. (2013); Hou et al. (2017); Kim et al. (2019); Orucevic et al. (2019); Baltres et al. (2020); Pawloski et al. (2021). A detailed comparison is given in Table B.3.

		Klein et al. (2013)	Hou et al. (2017)	Kim et al. (2019)	Orucevic et al. (2019)	Baltres et al. (2020)	Pawloski et al. (2021)	Al Masry et al. (2023)
Patients	(n_{train}, n_{test})	(817, 255)	(-, 163)	(208,76)	(65,754, 18,585)	(152, 168)	(2,587, 1,293)	(333, OOB)
Age	Mean	–	58.6	–	–	–	–	56.9
	Median	–	–	44.0	58	57.5	62	58.0
	Range	–	34-82	–	19-90	30-84	56-69	30-84
Clinico-pathologic features used for modeling	Tumor size	✓	✓	✓	✓	✓	✓	✓
	Tumor grade	✓	✓	✓	✓	✓	✓	✓
	Lymphovascular invasion			✓			✓	
	Lymph node status			✓		✓		✓
	ER	✓	✓	✓	✓	✓		✓
	PR	✓	✓	✓	✓	✓	✓	✓
	Ki67	✓	✓	✓		✓		✓
	p53							✓
ODX Prediction	Type	Continuous	Continuous	Classification	Classification	Classification	Classification	Distributional
	Threshold	< 18 $18 - 30$ > 30	< 18 $18 - 30$ > 30	< 11 > 25	≤ 25 > 25	< 18 $18 - 30$ > 30	≤ 25 > 25	≤ 25 > 25
Method		Multiple Linear Regression	Multiple Linear Regression	Neural Network Decision Jungle	Binomial Logistic Regression	Deep Multi-Layer Perceptron	Random Forest	Distributional Random Forest
Precision	Low risk	62.5-69.4%	72.6%	100%	87.5%	58.3%	92.9%	82.5%
	High risk	68.8-77.8%	–	25.0%	79.6%	63.0%	65.1%	62.3%
Sensitivity		58.6-59.1%	85.7%	11.0%	99.2%	55%	96.3%	92.0%
Specificity		70.5-77.4%	41.4%	100%	18.3%	78%	48.3%	40.2%
AUC		–	–	0.744	0.81	0.63	–	0.759

Table B.3: Comparison of our study with six selected published studies (Klein et al., 2013; Hou et al., 2017; Kim et al., 2019; Orucevic et al., 2019; Baltres et al., 2020; Pawloski et al., 2021) to predict the ODX score. For three classes only the sensitivity and specificity of the lower class are given.

B.4 Discussion

ODX is the most commonly available breast genomic test used in early-stage ER-positive/HER2-negative breast cancer. It makes it possible to define patients who are unlikely to benefit from chemotherapy. The ODX score is based on 6 gene groups. These groups correspond to the analysis of the markers in pathological reports. Some have compared the ODX score to this immuno-histological data and proved the predictive relationship with the ODX score. Several studies were published using this clinicopathological data to predict the ODX score with different methods (see Table B.3). The present study was realized to predict the ODX score from a specific regional cohort of 333 patients with clinical and immuno-histological data using distributional random forests. This prediction is associated with a predictive error on the one hand, and the ability to determine similar patients on the other hand. The proposed DRF model detected 82% of lower risk patients (ODX ≤ 25) and 62.3% of high risk patients (ODX > 25).

A few studies have proposed some prediction tools for the ODX score (Klein et al., 2013;Hou et al., 2017; Kim et al., 2019; Orucevic et al., 2019; Baltres et al., 2020; Pawloski et al., 2021). Each study is based on the specific categorization of patients according to the original ODX categories and TAILORx (see ODX Prediction Threshold in Table B.3). The prediction results of the different studies are similar and based on clinico-pathological data. The tumor size, tumor grade and PR are used in all the six selected published studies as well as for our current study. The Ki67 is not used in Orucevic et al. (2019) and Pawloski et al. (2021). In our study, we integrated the p53. The threshold used for the ODX score is different from one study to another. Our DRF model performs as well as the other prediction tools. The novelty is in providing additional information to the prediction (see Figure B.2) such as the probability of classes (low and high risk), the similar profiles and the uncertainty prediction.

The correct predicted values are 82.5% and 62.3% for low and high risk, respectively. We used the CRPS score to distinguish the best and worst predictions. The best results were obtained for ODX profiles below 16. The average Ki67, for the first best ten results, is under 14%, which corresponds to the low-risk profile of our previous study (Baltres et al., 2020). The average percentages of ER and PR are 93% and 77%, respectively, which fits into the same low-risk profiles. When looking at the surrounding family and the profile of close patients, we observe that similar profiles vary between 0 and 25 for ODX. The similarities fall in the low-risk profile. The Ki67 scores of similar profiles for the first ten results are below 25%.

As for the discordant results, they lie in the high-risk class. The averages of the ODX score, Ki67 and PR are respectively 46%, 36% and 22%. In addition, a negative correlation between ODX and PR for the best and worst results can be observed. The similar profiles for such cases have a high PR. This behavior is due to the small number of cases in the high-risk category. An example of the worst prediction is a patient with a high ODX score and probability of lying in the high class of 50%. The real ODX score is 49 and the predicted ODX score is near to the cut-off. The average ODX score for the 10 first similar profiles is 31 and the distribution is centered around 25. The similar profiles are very dispersed, which is difficult to analyze. Most of the nearest neighbors have an SBR grade of 3. The prediction is bad, but nevertheless, the similar profiles have a low ODX score and a high SBR grade. The size of the cohort and the training and testing phase could impact the prediction results. In addition, we have an unbalanced cohort in our study, since we have fewer patients in the high-risk class. In that case, the factors of the similar profiles that influence the ODX score such as PR, Ki67 and p53 should be considered. The distribution of identical profiles allows the clinician to retrieve similar historical cases in terms of evolution. The proposed model can be applied even when there is missing data. It makes it possible to predict the low-risk class with high certitude, which means no chemotherapy to plan. Our study is related to the dataset and it is therefore difficult to generalize to a different cohort because of known inter-cohort variability, especially on some biomarkers such as Ki67.

B.5 Conclusion

This paper proposes a new methodology for oncotype scoring prediction. This methodology is based on distributional random forests and uses 9 clinico-pathological features. It makes it possible to predict the distribution of the ODX score for a patient and provides an explanation of the predicted score by computing the probability of belonging to the low- or high-risk category and identifying the nearest similar profiles. The proposed Distributional Random Forest model detects 82% of low-risk patients (≤ 25) and 62.3% of patients with high risk (> 25). However, DRF presents certain limitations. The use of DRF with the pathologist's expertise on the different histological and immunohistochemical characteristics has a clinical impact to help oncologists in decision-making regarding breast cancer therapy. The medico-economic interest of this strategy is obvious. Additional studies are needed to further validate the DRF method and improve knowledge extraction from pathological data.

Appendix C

Additional comments

This appendix addresses three comments related to the work of Chapter 2 (Pic et al., 2023). Chapter 2 and its summary in the Introduction (Section 1.2) are prerequisites to the following comments.

C.1 Can the results of Chapter 2 be adapted to the logarithmic score?

During my follow-up committee, Christopher Ferro from the University of Exeter raised the question of the adaptation of the results in Chapter 2 (Pic et al., 2023) to other scoring rules, such as the logarithmic score, instead of the CRPS. Both the convergence rates and the upper bounds obtained are stated in a univariate setting and for a risk defined in terms of CRPS.

As introduced in Chapter 4, the logarithmic score (also known as ignorance score; Good 1952; Roulston and Smith 2002) is defined as

$$\mathrm{LogS}(F, y) = -\log(f(y)),$$

for y such that $f(y) > 0$ and where f is the probability density function (pdf) of the probabilistic forecast F.

The definition of convergence relies on the excess risk of the algorithm $\hat{F}_n$ (1.2) where the divergence of the scoring rule appears. When S is the logarithmic score, the divergence of S is the Kullback-Leibler divergence (Kullback and Leibler, 1951) :

$$\mathrm{div}_{\mathrm{LogS}}(F, G) = D_{KL}(G||F) = \int_{\mathbb{R}} \log\left(\frac{g(y)}{f(y)}\right) g(y)\mathrm{d}y$$

where f and g are the pdfs of F and G, respectively.

The Kullback-Leibler divergence is defined if G is absolutely continuous with respect to F, noted $G \ll F$. Thorarinsdottir and Schuhen (2018) explains this in simpler words: "The Kullback-Leibler divergence becomes ill-defined if the forecast distribution F has positive mass anywhere where the observation distribution G has mass zero." However, the proofs of the results of Chapter 2 rely on the k-nearest-neighbor (k-NN) algorithm (1.5) and the uniform kernel algorithm (1.6). These algorithms have atoms on $\{Y_{i:n}(X), 1 \leq i \leq k_n\}$ and $\{Y_i, \text{ s.t. } \|X_i - X\| \leq h_n\}$, respectively. This prevents F^* from being absolutely continuous with respect to them (at finite k_n and n).

Hence, the adaptation of convergence rates and upper bounds in Chapter 2 is not straightforward since the proof relies on the k-NN and uniform kernel algorithms. A potential solution

to circumvent the issue of absolute continuity could be to use kernel smoothing techniques (see, e.g., Wilks 2011, Section 3.3.6) to obtain densities. Kernel smoothing techniques are related to the postprocessing techniques known as ensemble dressing (see, e.g., Wilks 2018, Section 3.5.3).

C.2 Details on how the results of Chapter 2 adapt to the weighted CRPS

In the discussion of Chapter 2 (Section 2.4), it is mentioned that the results "*can easily be extended to the weighted CRPS*".

First, it should be specified that the *weighted CRPS* the article references is the threshold-weighted CRPS (twCRPS; Gneiting and Ranjan 2011). The threshold-weighted CRPS is defined as

$$\text{twCRPS}(F, y) = \int_{\mathbb{R}} w(z) \left(F(z) - \mathbb{1}_{y \leq z}\right)^2 \mathrm{d}z,$$

with w the weight function such that $w(z) \geq 0$. As mentioned in Thorarinsdottir and Schuhen (2018), the twCRPS reduces to the CRPS for $w(z) = 1$ and to the Brier score (Brier, 1950) for $w(z) = \mathbb{1}_{z=t}$. The strict propriety of the twCRPS is ensured if and only if the weight function w is strictly positive and integrable over $\mathbb{R}$. The choice of w allows the emphasis on regions of interest and eases the interpretability of forecast comparisons. In particular, it is used to emphasize the upper tail of the distributions when extreme events are of interest.

The two other families of weighted CRPS leading to proper scoring rules are the outcome-weighted CRPS (owCRPS; Holzmann and Klar 2017) and the vertically-rescaled CRPS (vrCRPS; Allen et al. 2023b). Gneiting and Ranjan (2011) also introduced a quantile-weighted CRPS where the weighting intervened in the expression of the CRPS as an integral of quantile scores. The results around the convergence in terms of CRPS can indeed easily be extended to its threshold-weighted counterpart and we provide more information on how the results can be adapted. However, whether the results can be extended to the owCRPS or the vrCRPS remains an open question and can lead to future research. Allen et al. (2023a) presents the three types of weighted CRPS and provides a comprehensive comparison of the different weighting methods.

Even if the results of Chapter 2 can be "*easily*" be adapted to the threshold-weighted CRPS, we want to explicit the details of this adaptation. As mentioned in the previous comment, both the convergence rates and upper bounds rely on the divergence and the entropy of the new scoring rule considered :

$$\begin{aligned}
\text{div}_{\text{twCRPS}}(F, G) &= \int_{\mathbb{R}} w(z)(F(z) - G(z))^2 \mathrm{d}z = \|F - G\|^2_{wL^2}; \\
\text{ent}_{\text{twCRPS}}(F) &= \int_{\mathbb{R}} w(z)F(z)(1 - F(z))\mathrm{d}z.
\end{aligned}$$

The divergence of the twCRPS is the weighted version of the squared L^2-norm between the cdfs.

Given a weight function w such that the twCRPS is strictly proper, the steps toward adapting the results are the following :

1. **Modification of the class of distributions $\mathcal{D}^{(h,C,M)}$**
   This is done by updating the following conditions of Definition 2.2 :

   ii) $\forall x \in [0,1]^d$, $\int_{\mathbb{R}} w(z)F^*_x(z)(1 - F^*_x(z))\mathrm{d}z \leq M$;
   iii) $\|F^*_{x'} - F^*_x\|_{wL^2} \leq C\|x' - x\|^h$, $\forall x, x' \in [0,1]^d$.

2. **Adapting the proofs of Propositions 2.1 and 2.2**
   Since the proofs rely on the integration of an upper bound of $\mathbb{E}[|\hat{F}_{n,x}(z) - F_x^*(z)|^2]$ (at fixed $x \in [0,1]^d$) with respect to z to obtain an upper bound of $\mathbb{E}[R_P(\hat{F}_n) - R_P(F^*)]$, the weight function w appears naturally and simply in the proofs.

3. **Adapting the proof of Proposition 2.3**
   This proof uses Lemmas 2.1 and 2.2 as well as the result of Problem 3.3 in Györfi et al. (2002) and considers the subclass where $Y \in \{0, L\}$. The adaptation of Lemma 2.1 requires m to be redefined as

$$m = \frac{\int_0^L w(z)(1 - F(z))\mathrm{d}z}{\int_0^L w(z)\mathrm{d}z},$$

   which coincides with the original definition for $w(z) = 1$. The adaptation of Lemma 2.2 leads to

$$\mathrm{twCRPS}(F, y) = W(L)\mathrm{Brier}(p, \frac{y}{L}),$$

   where $W(L) = \int_0^L w(z)\mathrm{d}z$ and where F and y are a binary forecast and a binary observation, respectively, taking values in $\{0, L\}$.

This adaptation leads to the same optimal minimax rates of convergence on the adapted class of distribution $\mathcal{D}^{(h,C,M)}$ and to the same upper bounds. However, the constants (h, C, M) may depend on the weight function w.

C.3 Analog ensemble techniques, k-nearest-neighbor algorithm and Cover-Hart inequality

Recall that given a training sample $\mathcal{D}_n = \{(X_i, Y_i), 1 \leq i \leq n\}$, the k-nearest-neighbor (k-NN) and uniform kernel algorithms give the following cdf estimators :

$$\hat{F}_{n,X}(z) = \frac{1}{k}\sum_{i=1}^{k} \mathbb{1}_{Y_{i:n}(X) \leq z}; \qquad (k\text{-NN})$$

$$\hat{F}_{n,X}(z) = \frac{\sum_{i=1}^n \mathbb{1}_{\|X - X_i\| \leq h}\mathbb{1}_{Y_i \leq z}}{\sum_{i=1}^n \mathbb{1}_{\|X - X_i\| \leq h}}, \qquad (\text{uniform kernel})$$

with $X \in \mathbb{R}^d$ the covariates associated with the prediction and $z \in \mathbb{R}$. For a given X, the bandwidth $h = h(X)$ can be chosen such that the number of $(X_i)_{1 \leq i \leq n}$ is equal to k. Vice-versa, for a given X, the number of neighbors $k = k(X)$ can be chosen such that the number of neighbors coincides with the number of $(X_i)_{1 \leq i \leq n}$ within the bandwidth h. Issues might appear when two covariates in the training sample are at the same distance of X but this case is marginal. Even though the equivalency strictly holds only for a fixed training sample and requires the hyperparameters (i.e., k or h) to depend on X, k-NN and uniform kernel algorithms are strongly related. They only differ in their definitions of neighborhood, either by restricting the number of neighbors or the span of the neighborhood.

In statistical postprocessing of weather forecasts, analog ensemble methods are nonparametric methods estimating a probabilistic distribution with an ensemble of past observations that are similar, called *analogs*, to the current state of the atmosphere. In order to obtain analogs that are similar enough to lead to an informative analog ensemble, most methods treat the problem as local, such that analogs are selected for each location and forecast lead time (Hamill and Whitaker, 2006). The search for analogs depends on the size of the historical dataset used

as training data and on the dimension of the search space (i.e., the number of features compared and their range). A local approach based on selected features, as well as a large training sample, improves the pertinence of the analog ensembles. With this general definition, analog ensemble methods encapsulate numerous statistical learning techniques: k-NN when the ensemble size needs to be fixed, uniform kernel methods when the span of the neighborhood of analogs needs to be controlled, kernel methods when the weights of the analogs in not uniform or when the span of the neighborhood is not limited, and random forests (e.g., quantile regression forests; Taillardat et al. 2016) when a more complex relation characterizes analogs.

In general, analog ensemble methods mainly differ on the metric or similarity measure they rely on and on the features selected to define analogs. In order to perform the postprocessing of ensemble forecasts, the training sample is composed of pairs of raw output of the ensemble forecast and corresponding observations. In particular, only summary statistics of the raw ensemble forecast and most informative variables are kept as selected features to avoid having a search space too large. Standard metrics, such as the Euclidean distance, have been used but more complex metrics, such as the one proposed in Delle Monache et al. (2013), have outperformed simpler metrics. The metric provided in Delle Monache et al. (2013) considers the features' standard deviation and allows the analog ensembles to capture the flow-dependent forecast uncertainty. The metric is not the only means to capture the flow-dependent uncertainty: k-NN and uniform kernel algorithm can capture flow dependency by allowing their hyperparameters to depend on the features, i.e. allowing the number of analogs $k = k(X)$ or the span of the analogs $h = h(X)$ to vary with the flow. The importance of flow-dependent uncertainty in weather forecasting has been introduced in Section 1.1.1.

In classical regression, an estimate based on infinite samples can, at best, halve the theoretical risk of an estimate based on one sample. This result holds under a large class of loss functions, called the Cover-Hart family. Gneiting (2012) showed that this result, known as the Cover-Hart inequality (Cover and Hart, 1967), can be adapted to distributional regression for the class of kernel scores. Let $\mathcal{P}$ be the family of the Radon probability measures on a Hausdorff space $(\Omega, \mathcal{B})$, where $\mathcal{B}$ is the Borel-σ-algebra. Let $\mathrm{S} : \mathcal{P} \times \Omega \to \mathbb{R}$ be a kernel score, then

$$\mathbb{E}_P[S(P, Y')] \leq \mathbb{E}_P[S(\delta_Y, Y')] \leq 2\mathbb{E}_P[S(P, Y')] \tag{C.1}$$

for all probability measures $P \in \mathcal{P}$, where Y and Y' are independent with distribution P. The class of kernel scores encapsulates the CRPS (Matheson and Winkler, 1976) and the Brier score (Brier, 1950) in the univariate case, as well as the energy score (Gneiting and Raftery, 2007) and the variogram score (Scheuerer and Hamill, 2015b) in the multivariate case. The second inequality in (C.1) is an equality but is showcased in that way to explicit the link with the Cover-Hart inequality in point regression. This result is stated in the unconditional framework but can be extended to the conditional framework that is of interest in distributional regression.

The Cover-Hart inequality (C.1) can be adapted to a conditional setting and provide insights on distributional regression. In the conditional framework, the one-sample estimator is the 1-NN estimator (i.e., an analog ensemble method where only the closest analog is used). This implies that, in distributional regression, the 1-NN estimator has a risk that is exactly twice the Bayes risk (i.e., the minimal risk which is associated with the true distribution of Y given X) when the risk is associated with a kernel score.

This result is somehow encouraging because the best improvement (of a factor two) over the 1-NN estimator can be obtained in distributional regression. Even though this result applies to the same framework as the results of Chapter 2, this is the only meaningful connection that exists between these two works.

Appendix D

Résumé long

La précision des prévisions météorologiques est cruciale dans divers domaines (par exemple, les énergies renouvelables, les réseaux de transport ou l'agriculture), tant pour la prise de décision que pour son impact financier (Palmer, 2002). Les prévisions probabilistes sont une composante essentielle de la prise de décision optimale car elles quantifient l'incertitude de la prévision (Gneiting and Katzfuss, 2014). Dans les prévisions météorologiques, le post-traitement statistique est nécessaire pour produire des prévisions probabilistes calibrées et précises à partir de systèmes de prévision d'ensemble. Cette thèse se concentre sur trois aspects différents du post-traitement statistique : les taux de convergence théoriques, le post-traitement des précipitations sur grille et la vérification des prévisions probabilistes spatiales.

Cette thèse a été réalisée en collaboration avec Météo-France sous la supervision directe de Maxime Taillardat mais aussi par des échanges avec d'autres membres de Météo-France et l'utilisation de leurs moyens de calcul haute performance.

Les méthodes de post-traitement statistique utilisent la sortie d'un modèle physique pour améliorer la prédiction d'une variable d'intérêt. Les règles de sore sont utilisées pour la vérification probabiliste des prévisions afin de mesurer et de comparer la performance prédictive de prévisions concurrentes. Cette thèse étudie différents aspects du post-traitement statistique et de la vérification des prévisions probabilistes.

- D'un point de vue théorique, seuls des résultats limités sont disponibles concernant la convergence des méthodes de post-traitement. Chapitre 2 (Pic et al., 2023) est une contribution théorique qui se concentre sur le taux de convergence optimal minimax pour le risque théorique associé au *continuous ranked probability score*.

- Dans les applications de prévisions météorologiques, les prévisions spatiales sont omniprésentes. Cependant, les méthodes basées sur les forêts aléatoires qui sont utilisées de manière opérationnelle pour post-traiter les prévisions à Météo-France ne prennent pas vraiment en compte le cadre spatial. De plus, elles souffrent de la voracité de la mémoire de stockage et d'une incapacité à extrapoler. Dans le Chapitre 3 (Pic et al., 2024b), nous proposons une méthode de régression distributionnelle basée sur les U-Nets pour post-traiter les ensembles en contournant la voracité de la mémoire de stockage tout en obtenant une performance prédictive comparable aux méthodes état de l'art.

- En ce qui concerne la vérification des prévisions probabilistes, aucune règle de score unique n'est en mesure de fournir une évaluation idéale de la performance prédictive des prévisions et, par conséquent, différentes règles de score devraient être utilisées pour la comprendre. Cette affirmation est d'autant plus importante dans le cadre de la vérification des prévisions spatiales que la performance prédictive est soumise à des caractéristiques complexes. Dans cette optique, les règles de score interprétables sont des outils puissants qui facilitent la vérification des prévisions. Regarding probabilistic forecast verification, no single règle de score is able to provide an ideal assessment of the predictive performance of forecasts, and thus, different règles de score should be used to understand it. This statement is even more important in a spatial forecast verification setting as predictive performance is subject to complex characteristics. With that in mind, interpretable règles de score are powerful tools facilitating forecast verification. Le Chapitre 4 (Pic et al., 2024a) présente la manière dont les principes d'agrégation et de transformation peuvent être utilisés pour construire des règles de score multivariées interprétables.

De plus, le Chapitre 5 fournit des perspectives sur les travaux qui composent cette thèse de doctorat.

Le reste de ce chapitre est organisé comme suit. La Section D.1 introduit le contexte nécessaire à la compréhension de la contribution de cette thèse. Les sections D.2, D.3 et D.4

résument les travaux liés aux Chapitres 2, 3 et 4, respectivement. La Section D.5 résume brièvement les travaux présentés dans les annexes A et B.

D.1 Introduction générale

D.1.1 Incertitude dans la modélisation de systèmes déterministes et prévisions d'ensemble

Une approche intuitive de la prévision météorologique consiste à considérer que la physique de l'atmosphère est régie par un ensemble d'équations différentielles non-linéaires déterministes (Bjerknes, 2009). Cependant, dans les années 1960, Lorenz (1963) a montré que l'atmosphère est un système chaotique caractérisé par de multiples sources d'incertitude (voir Wilks and Vannitsem 2018 pour plus de détails).

La sensibilité aux conditions initiales combinée à l'incertitude des conditions initiales représente une source majeure d'incertitude pour les prévisions météorologiques. L'incertitude des conditions initiales découle de divers aspects, tels que la combinaison de différents types d'observations et la variation de la qualité et de la couverture en fonction de la variable d'intérêt, de l'emplacement et de la méthode de mesure. Par exemple, comme indiqué dans le Chapitre 3, la qualité des mesures radar des précipitations dépend de la distance par rapport à l'instrument et de l'orographie sous-jacente (voir, par exemple, Germann et al. 2022). Le domaine de l'assimilation des données est consacré à la combinaison de différentes sources de données afin de fournir des conditions initiales bien adaptées aux systèmes de prévision numérique du temps (PNT). En outre, dans la pratique, les modèles dynamiques de prévision météorologique ne décrivent pas parfaitement la dynamique réelle. Tout d'abord, le modèle peut fournir une modélisation incorrecte des phénomènes en jeu. Deuxièmement, tous les systèmes de prévision météorologique numérique sont incomplets en raison de la discrétisation spatiale et temporelle et de la paramétrisation de processus physiques non résolus.

De plus, l'atmosphère a une prévisibilité dépendante du flux, ce qui signifie que la propagation de l'incertitude de la condition initiale dépend de l'état du système. Les erreurs de prévision fluctuent donc d'un bout à l'autre du globe et en fonction de la variable considérée, mais aussi d'un jour à l'autre (Buizza, 2018). Ces limites affectent également d'autres modèles physiques tels que les modèles climatiques et les modèles hydrologiques.

Des prévisions d'ensemble ont été élaborées pour tenter de tenir compte des incertitudes des modèles. Cependant, le choix d'un ensemble bien adapté est difficile, car l'échantillonnage aléatoire basé sur une gamme de résultats possibles ne conduit pas à un ensemble informatif. En outre, un grand nombre de membres peut être intéressant mais coûteux en termes de calcul, et une augmentation de la résolution est souvent préférée car elle permet de résoudre des processus à des échelles plus fines. Afin d'échantillonner un système comportant des millions de degrés de liberté avec quelques dizaines de membres, différentes approches se sont avérées capables de représenter les incertitudes du modèle : approches multimodèles, approches perturbées, approches à tendance perturbée, approches à rétrodiffusion stochastique, et des combinaisons de ces approches. Les lecteurs peuvent se référer à Buizza (2018) pour un aperçu historique de l'utilisation des prévisions d'ensemble.

Malgré l'amélioration continue des systèmes de prévision numérique au cours des dernières décennies (Bauer et al., 2015), l'amélioration des performances prédictives des variables proches de la surface est plus lente que celle des variables situées plus haut dans l'atmosphère (Buizza, 2018). Les prévisions d'ensemble émises par les systèmes de PNT souffrent de biais et de sous-dispersion. Ce phénomène affecte tous les systèmes de prévision numérique, indépendamment

du service météorologique et de la variable physique concernée. Comme les systèmes dynamiques sur lesquels elles reposent, les erreurs des prévisions d'ensemble varient en fonction de la variable d'intérêt et de la région concernée. En outre, l'augmentation de l'échéance (c'est-à-dire le temps écoulé entre l'émission d'une prévision et sa validité) est associée à une diminution de la prévisibilité. Comme ces erreurs sont systématiques, elles peuvent être corrigées par des approches statistiques appelées *méthodes statistiques de post-traitement.*

D.1.2 Post-traitement statistique

Les méthodes de post-traitement statistique visent à utiliser les paires passées de l'ensemble brut et d'observations pour améliorer la prévision d'une variable d'intérêt. Le terme *ensemble brut* désigne la sortie d'ensemble non traitée (c'est-à-dire brute) des systèmes de PNT. L'objectif étant de fournir des prévisions informatives aux utilisateurs finaux, les prévisions doivent être probabilistes. *Prévisions probabilistes* fournissent une prévision sous la forme d'une distribution. Cela permet de quantifier l'incertitude de la prévision, garantissant ainsi une prise de décision optimale (Gneiting and Katzfuss, 2014). L'ensemble brut est une prévision probabiliste car il peut être interprété comme une distribution empirique où tous les membres ont la même probabilité. Les prévisions probabilistes peuvent prendre la forme de n'importe quelle formulation capable de décrire l'intégralité de la distribution de probabilité. Dans un cadre univarié, elles peuvent prendre la forme d'une fonction de densité de probabilité, d'une fonction de répartition ou d'une fonction quantile, par exemple.

Les méthodes de post-traitement statistique peuvent être classées de plusieurs manières. Nous présentons trois classifications différentes de méthodes basées sur leur paramétrage distributionnel, leur utilisation et leur complexité. Premièrement, les méthodes de post-traitement statistique peuvent être classées en deux groupes ("non paramétriques" et "paramétriques") sur la base de l'hypothèse d'une famille de distributions. Les méthodes non paramétriques comprennent *analog ensemble* (voir, par exemple, Delle Monache et al. 2013) qui utilise des situations atmosphériques antérieures similaires pour améliorer l'ensemble brut. *Analog ensemble* est lié aux méthodes de k-nearest neighbor (k-NN; k plus proches voisins), comme expliqué dans l'Annexe C. *Quantile regression forest* (QRF ; Taillardat et al. 2016) est une méthode non paramétrique qui utilise les données dans les nœuds terminaux (c'est-à-dire les feuilles) d'une forêt aléatoire pour calculer une moyenne pondérée des distributions empiriques. Les méthodes paramétriques comprennent l'*ensemble model output statistics* (EMOS ; Gneiting et al. 2005), qui suppose que la distribution prédite est une distribution normale dont les paramètres dépendent linéairement des statistiques sommaires de l'ensemble brut. Les méthodes paramétriques fournissent une famille de distributions paramétriques adaptées à la variable considérée (par exemple, basées sur la théorie des valeurs extrêmes ; Friederichs et al. 2018). La plupart des méthodes non paramétriques n'ont pas de capacité d'extrapolation au-delà de la plage des données observées, mais sont capables de conserver les caractéristiques de la vraie distribution à partir des données observées. La frontière entre les deux classes est poreuse : les QRF avec extension de queue (TQRF ; Taillardat et al. 2019) sont une méthode semi-paramétrique qui ajuste une distribution paramétrique sur la sortie d'une QRF. La classification en méthodes paramétriques et non paramétriques est examinée plus en détail dans Vannitsem et al. (2021).

Deuxièmement, les méthodes de post-traitement statistique diffèrent également dans leur utilisation. L'utilisation la plus courante du post-traitement consiste à post-traiter séparément les marginales univariées et la structure de dépendance. La structure de dépendance peut être obtenue à partir de l'ensemble brut, comme dans le cas du couplage de copules d'ensemble (ECC ; Schefzik et al. 2013), ou à partir d'observations historiques, comme dans le cas *Schaake shuffle* (ScS ; Clark et al. 2004). Par ailleurs, si l'ensemble brut ou les données historiques ne modélisent pas suffisamment bien la structure de dépendance, ils peuvent être post-traités à

l'aide de techniques adaptées telles qu'une approche par copule gaussienne (voir, par exemple, Möller et al. 2013). Certaines méthodes de post-traitement statistique prennent directement en compte des quantités multivariées (par exemple, Pinson et al. 2009). Les lecteurs peuvent se référer à Schefzik and Möller (2018) pour plus de détails sur le post-traitement de la structure de dépendance. Certaines méthodes traitent directement et simultanément chaque membre de l'ensemble brut. *Member-by-member* (Van Schaeybroeck and Vannitsem, 2015) corrige la moyenne et la dispersion de l'ensemble via une combinaison linéaire des prédicteurs bruts. Le post-traitement des ensembles à l'aide de transformers (PoET ; Ben Bouallègue et al. 2024b) utilise des transformateurs dans une architecture U-Net pour post-traiter les membres de l'ensemble.

Troisièmement, les méthodes de post-traitement statistique diffèrent par leur niveau de complexité. Les méthodes les moins complexes sont liées aux méthodes d'apprentissage statistique, comme par exemple *analog ensemble* et k-NN et QRF et les forêts aléatoires (Breiman, 2001; Meinshausen, 2006). La relative simplicité de ces méthodes permet plus de simplicité mais moins de flexibilité en termes de modélisation de la dépendance d'une variable d'intérêt en fonction de prédicteurs. Des méthodes plus complexes issues du machine learning peuvent également être employées. Les *distributional regression networks* (DRN ; Rasp and Lerch 2018) sont une approche basée sur les réseaux neuronaux (NN) qui prédit les paramètres d'une distribution d'intérêt. Elle tire parti de la flexibilité des réseaux de neurones entièrement connectés pour modéliser la dépendance des paramètres par rapport aux covariables (utilisées en entrée du DRN). Le DRN peut être considéré comme une extension d'EMOS. Au lieu de modéliser linéairement la dépendance des paramètres sur les statistiques sommaires de l'ensemble brut, il permet de prendre en compte des dépendances non linéaires plus souples. À l'extrémité supérieure du spectre de complexité se trouvent les méthodes basées sur les techniques de deep learning (DL). L'approche PoET, présentée ci-dessus, utilise des transformers qui ont été introduits à l'origine pour les tâches de traitement du langage naturel (Vaswani et al., 2017). Toutes les méthodes basées sur l'apprentissage profond ne représentent pas le même niveau de complexité. La complexité peut s'accompagner d'une augmentation de la flexibilité, mais aussi de la difficulté de mise en œuvre. Le Chapitre 3 propose une méthode basée sur U-Net pour prédire les distributions paramétriques qui étend DRN aux données en grille.

Ces trois classifications donnent un premier aperçu du large éventail de méthodes de post-traitement statistique. Pour des aperçus plus détaillés, les lecteurs peuvent se référer à Taillardat et al. (2019), Vannitsem et al. (2021) et Schulz and Lerch (2022b).

Comme nous l'avons brièvement mentionné, différentes méthodes de post-traitement peuvent être préférées en fonction de l'application. Nous nous concentrons brièvement sur les méthodes paramétriques pour discuter explicitement de la manière dont les variables d'intérêt diffèrent en termes de post-traitement. Tout d'abord, différentes familles de distribution sont adaptées à différentes variables. Par exemple, la température et la pression au niveau de la mer peuvent être modélisées par des distributions normales. D'autres variables peuvent présenter des distributions asymétriques, multimodales ou discontinues qui peuvent être modélisées à l'aide de familles de distribution tronquées, censurées ou mixtes. Par exemple, les précipitations présentent une masse atomique en zéro liée aux événements secs (c'est-à-dire l'absence de précipitations). En outre, les précipitations présentent souvent une queue lourde ; une famille de distribution s'appuyant sur la théorie des valeurs extrêmes peut donc améliorer le post-traitement. Lerch and Thorarinsdottir (2013) a proposé une variante d'EMOS utilisant une distribution (GEV) pour le post-traitement de la vitesse maximale quotidienne du vent. Taillardat et al. (2019) présente TQRF comme une extension de QRF pour améliorer la prévision des précipitations extrêmes. Un examen complet du post-traitement des événements extrêmes est fourni dans Friederichs et al. (2018).

En outre, Hemri et al. (2014) et Taillardat and Mestre (2020) ont souligné que toutes

les variables d'intérêt ne représentent pas la même difficulté en termes de post-traitement. Les variables présentant une dépendance spatio-temporelle à courte échelle (par exemple, les précipitations ou les rafales de vent) sont plus difficiles à traiter que les variables lisses dans l'espace (par exemple, la température de surface ou la pression au niveau de la mer). Dans le même ordre d'idées, Schulz and Lerch (2022b) indique que "les rafales de vent sont une variable cible météorologique difficile car elles sont régies par des processus à petite échelle et une occurrence locale, de sorte que leur prévisibilité est limitée même pour les modèles de prévision météorologique numérique exécutés à des résolutions permettant la convection". La prévisibilité des variables est liée à leurs caractéristiques physiques et à leur représentation dans les modèles de PNT.

D.1.3 Vérification de prévisions probabilistes

La vérification des prévisions probabilistes répond à deux objectifs principaux : quantifier la performance prédictive d'une prévision et comparer des prévisions concurrentes. Dans le contexte du post-traitement statistique, la prévision de référence évidente est l'ensemble brut et les techniques de post-traitement devraient améliorer la performance prédictive par rapport à cette référence.

Gneiting et al. (2007) a proposé un paradigme pour l'évaluation des prévisions probabilistes : "maximizing the sharpness of the predictive distributions subject to calibration." La *calibration* est la compatibilité statistique entre les prévisions et les observations. La *sharpness* est la concentration de la prévision et est une propriété de la prévision elle-même. En d'autres termes, le paradigme vise à minimiser l'incertitude de la prévision, étant donné que la prévision est statistiquement cohérente avec les observations. Ce principe d'évaluation des prévisions probabilistes a fait l'objet d'un consensus dans le domaine des prévisions probabilistes (voir, par exemple, Gneiting and Katzfuss 2014; Thorarinsdottir and Schuhen 2018).

Pour les prévisions univariées, il existe plusieurs définitions de calibration en fonction du contexte. La définition la plus utilisée est la *calibration probabiliste* et, d'une manière générale, elle consiste à calculer le rang des observations parmi les échantillons de la prévision et à vérifier l'uniformité par rapport aux observations. Si la prévision est calibrée, les observations ne devraient pas pouvoir être distinguées des échantillons de la prévision et la distribution de leurs rangs devrait donc être uniforme. La calibration probabiliste peut être évaluée par des histogrammes de transformation intégrale de probabilité (Dawid, 1984) ou des histogrammes de rang (Anderson, 1996; Talagrand et al., 1997) pour les prévisions d'ensemble lorsque les observations sont stationnaires (c.-à-d. que leur distribution est la même dans le temps). Les lecteurs intéressés par une compréhension plus approfondie de l'étalonnage des prévisions univariées sont invités à consulter Tsyplakov (2013, 2020). Pour les prévisions multivariées, une approche populaire repose sur un principe similaire : tout d'abord, les échantillons de prévisions multivariées sont transformés en quantités univariées à l'aide de ce que l'on appelle des fonctions de pre-rank, puis la calibration est évaluée à l'aide de techniques utilisées dans le cas univarié (voir, par exemple, Gneiting et al. 2008; Allen et al. 2024).

D'un point de vue quantitatif, les règles de score fournissent une évaluation quantitative de la qualité d'une prévision probabiliste au regard de l'observation qui se concrétise. Une règle de score S attribue une quantité réelle $\mathrm{S}(F, y)$ à une paire prévision-observation (F, y), où $F \in \mathcal{F}$ est une prévision probabiliste et $\boldsymbol{y} \in \mathbb{R}^m$ une observation. Dans la convention orientée négativement, une règle de score S est *propre par rapport à la classe* $\mathcal{F}$ si

$$\mathbb{E}_G[\mathrm{S}(G, \boldsymbol{Y})] \leq \mathbb{E}_G[\mathrm{S}(F, \boldsymbol{Y})] \tag{D.1}$$

pour tout $F, G \in \mathcal{F}$, où $\mathbb{E}_G[\cdots]$ est l'espérance par rapport à $\boldsymbol{Y} \sim G$. En termes simples, une règle de score est propre par rapport à une classe de distribution si sa valeur attendue

est minimale lorsque la vraie distribution est prédite, pour n'importe quelle distribution de la classe. En outre, la règle de score S est *strictement propre par rapport à la classe* $\mathcal{F}$ si l'égalité dans (D.1) se vérifie si et seulement si $F = G$. Cela garantit la caractérisation de la prévision idéale (c'est-à-dire qu'il existe une prévision unique associée à l'espérance minimale et qu'il s'agit de la vraie distribution). En outre, les règles de score propres sont des outils puissants car elles permettent d'évaluer simultanément la calibration et la sharpness (Winkler, 1977; Winkler et al., 1996).

Toutefois, comme le rappellé dans le Chapitre 4, la propreté (stricte) ne suffit pas pour obtenir des règles de score informatives. Nous proposons un cadre pour construire des règles de score propres interprétables qui sont plus informatives dans la vérification des prévisions probabilistes spatiales.

De plus, comme les méthodes de post-traitement statistique apprennent à prédire une distribution probabiliste sur la base d'observations passées, leur évaluation pratique devrait être basée sur un ensemble indépendant de données non vues afin d'éviter tout biais potentiel. Dans la pratique, des limitations supplémentaires peuvent résulter de la saisonnalité ou du manque de cohérence des données (par exemple, en raison du changement climatique ou des mises à jour des systèmes de prévision numérique du temps).

D.2 Distributional regression and its evaluation with the CRPS: bounds and convergence of the minimax risk

De nombreuses méthodes de post-traitement statistique reposent sur la régression distributionnelle. Les méthodes de post-traitement visent à modéliser la distribution conditionnelle d'une variable d'intérêt $Y \in \mathbb{R}^m$ (par exemple, les précipitations cumulées sur 3 heures) compte tenu de la sortie d'un modèle physique $X \in \mathbb{R}^d$ (par exemple, sous la forme de résumés statistiques), dénotée F_X^*. Dans un contexte de vérification, les règles de score sont utilisées pour mesurer et comparer les performances prédictives de prévisions probabilistes concurrentes. Les règles de score peuvent être considérées comme l'équivalent des fonctions de perte (également connues sous le nom de fonctions de score) dans la régression ponctuelle.

Soit $\overline{\mathrm{S}}(F, G) = \mathbb{E}_G[\mathrm{S}(F, Y)]$ le score espéré de F pour la règle de score S. En régression distributionnelle, la performance prédictive d'une prévision probabiliste $\hat{F} : x \mapsto \hat{F}_x$ est évaluée par son risque théorique

$$\begin{aligned} R_P(\hat{F}) &= \mathbb{E}_{(X,Y)\sim P}\left[\mathrm{S}(\hat{F}_X, Y)\right]; \\ &= \mathbb{E}_{X\sim P_X}\left[\overline{\mathrm{S}}(\hat{F}_X, F_X^*)\right], \end{aligned}$$

où P est la distribution jointe de (X, Y) et P_X est la distribution marginale de X. Si S est strictement propre, alors F^* est un prédicteur de Bayes et son risque théorique

$$\begin{aligned} R_P(F^*) &= \mathbb{E}_{(X,Y)\sim P}[\mathrm{S}(F_X^*, Y)]; \\ &= \mathbb{E}_{X\sim P_X}\left[\overline{\mathrm{S}}(F_X^*, F_X^*)\right] \end{aligned}$$

est le risque de Bayes. Nous rappelons que le risque de Bayes est le risque théorique minimal pour tous les prédicteurs possibles et qu'un prédicteur de Bayes est un prédicteur qui atteint le risque de Bayes. En outre, si S est strictement correct, l'ensemble des prédicteurs de Bayes sont les prévisions $\hat{F}$ telles que $\hat{F}_X = F_X^*$ P_X-presque partout.

Les techniques de post-traitement statistique reposent sur un échantillon d'apprentissage $D_n = \{(X_i, Y_i), 1 \leq i \leq n\}$ et sont évaluées en termes de performance prédictive par rapport à de nouvelles données (X, Y). L'échantillon d'apprentissage D_n et les données de test (X, Y)

sont indépendants et identiquement distribués à partir de la même distribution P. Étant donné l'échantillon d'apprentissage D_n, un algorithme $\hat{F}_n : x \mapsto \hat{F}_{n,x}$ est construit pour estimer la distribution conditionnelle F^*_x. Dans ce contexte, le risque théorique de $\hat{F}_n$ est exprimé comme suit

$$\begin{aligned}\mathbb{E}_{D_n\sim P^n}\left[R_P(\hat{F}_n)\right] &= \mathbb{E}_{D_n\sim P^n}\mathbb{E}_{(X,Y)\sim P}\left[\mathrm{S}(\hat{F}_{n,X}, Y)\right];\\ &= \mathbb{E}_{D_n\sim P^n}\mathbb{E}_{X\sim P_X}\left[\overline{\mathrm{S}}(\hat{F}_{n,X}, F^*_X)\right].\end{aligned}$$

Le risque théorique est moyenné sur les valeurs possibles de l'échantillon d'apprentissage D_n, ce qui le rend uniquement dépendant de la distribution P et de la taille de l'échantillon n. Comme précédemment, si la règle de score S est strictement propre, F^* est un prédicteur de Bayes et son risque est le risque de Bayes. La quantité d'intérêt est alors l'excès de risque défini comme la différence entre le risque théorique d'un algorithme $\hat{F}_n$ et le risque de Bayes :

$$\mathbb{E}_{D_n\sim P^n}\left[R_P(\hat{F}_n)\right] - R_P(F^*) = \mathbb{E}_{D_n\sim P^n}\mathbb{E}_{X\sim P_X}\left[\overline{\mathrm{S}}(\hat{F}_{n,X}, F^*_X) - \overline{\mathrm{S}}(F^*_X, F^*_X)\right]. \tag{D.2}$$

Lorsque S est propre, la différence entre les scores espérés du côté droit de l'équation est appelée divergence de S (voir, par exemple, Gneiting and Katzfuss 2014, Section 3.1 et Thorarinsdottir et al. 2013).

Nous nous intéressons aux résultats de convergence en régression distributionnelle. Le cadre présenté ci-dessus est largement utilisé dans la pratique mais manque de garanties théoriques. La plupart des énoncés de convergence dans la régression distributionnelle sont non seulement dérivés dans un cadre inconditionnel, mais supposent également des tailles d'échantillon arbitrairement grandes (voir, par exemple, Thorey et al. 2017 et Mösching and Dümbgen 2020). Une exception est la régression distributionnelle isotonique qui, sous des hypothèses de monotonicité, minimise le *continuous ranked probability score* (CRPS) et est consistante au sens de la distance de Kolmogorov (Henzi et al., 2021).

Nous nous concentrons sur le cas univarié ($m = 1$) car il correspond au cadre de nombreuses méthodes de post-traitement statistique. De plus, nous choisissons le CRPS (Matheson and Winkler, 1976) comme étant la règle de score qui nous intéresse, qui a l'avantage d'être strictement propre par rapport à $\mathcal{P}_1(\mathbb{R})$ (c.-à-d. les distributions sur $\mathbb{R}$ avec un premier moment fini). Puisque $m = 1$, les distributions sont identifiées à leur fonction de répartition (cdf). La divergence du CRPS est la norme L^2 de la différence entre la cdf de $\hat{F}_X$ et la cdf conditionnelle F^*_X (également connue sous le nom de distance de Cramér au carré du second ordre ou de distance quadratique intégrée; Thorarinsdottir et al. 2013).

En régression ponctuelle, il est nécessaire de restreindre la convergence sur une classe donnée de distribution pour obtenir des résultats non triviaux (Stone 1982; Györfi et al. 2002). Afin d'étudier le taux de convergence de l'excès de risque (D.2) vers zéro lorsque $n \to \infty$, nous introduisons la notion de *taux de convergence minimax optimal*. Le risque minimax correspond au meilleur risque réalisable dans le pire des cas (d'où le nom minimax). Plus précisément, étant donné une classe de distributions $\mathcal{D}$, le taux de convergence optimal minimax quantifie l'erreur minimale qu'un algorithme $\hat{F}_n$ peut atteindre uniformément sur une classe donnée de distributions $\mathcal{D}$, lorsque la taille n de l'ensemble d'apprentissage D_n devient importante. La définition formelle du taux de convergence minimax en régression distributionnelle est la suivante.

Definition D.1. *Une suite positive (a_n) est appelée taux de convergence optimal minimax sur la classe $\mathcal{D}$ si*

$$\liminf_{n\to\infty} \inf_{\hat{F}_n} \sup_{P\in\mathcal{D}} \frac{\mathbb{E}_{D_n\sim P^n}[R_P(\hat{F}_n)] - R_P(F^*)}{a_n} > 0 \tag{D.3}$$

et

$$\sup_{n\to\infty} \inf_{\hat{F}_n} \sup_{P\in\mathcal{D}} \frac{\mathbb{E}_{D_n\sim P^n}[R_P(\hat{F}_n)] - R_P(F^*)}{a_n} < \infty, \tag{D.4}$$

où l'infimum est pris sur tous les modèles de régression distributionnelle $\hat{F}_n$ formés sur D_n. Si la suite (a_n) ne satisfait que la borne inférieure (D.3)*, elle est appelée taux de convergence minimax minimal.*

Nous considérons les classes de distributions suivantes.

Definition D.2. *Pour $h \in (0,1]$, $C > 0$ et $M > 0$, soit $\mathcal{D}^{(h,C,M)}$ la classe de distributions P telle que $F_x^*(y) = P(Y \leq y|X = x)$ vérifie :*

i) $X \in [0,1]^d$ P_X*-a.s. ;*

ii) Pour tout $x \in [0,1]^d$, $\int_{\mathbb{R}} F_x^*(z)(1 - F_x^*(z))\mathrm{d}z \leq M$ *;*

iii) $\|F_{x'}^* - F_x^*\|_{L^2} \leq C\|x' - x\|^h$ *pour tous* $x, x' \in [0,1]^d$.

Les conditions définissant la classe de distribution $\mathcal{D}^{(h,C,M)}$ sont analogues aux conditions en régression ponctuelle. Nous proposons une interprétation et une discussion des conditions $i) - iii)$. La condition $i)$ est une condition sur les covariables et peut être étendue à un compact. Cette condition découle du fait que l'augmentation du nombre d'échantillons n tente de remplir l'espace des covariables afin d'avoir un échantillon d'apprentissage représentatif de toutes les valeurs possibles. Par conséquent, chaque point de covariable doit être accessible et l'étendue de l'espace des covariables a un impact sur la convergence. La condition $ii)$ limite la sharpness (ou l'entropie) de F_x^* pour tout $x \in [0,1]^d$. La sharpness est associée à l'information transportée par la distribution, et il peut sembler intuitif que moins il y a d'information transportée (c'est-à-dire plus M est grand), plus il faut d'échantillons pour obtenir la même performance prédictive (en termes de risque théorique). La condition $iii)$ est une condition de régularité imposant que des covariables proches conduisent à des distributions conditionnelles proches. Puisque l'algorithme $\hat{F}_n$ utilise la connaissance des observations passées pour estimer la distribution conditionnelle à $X = x$ et que l'augmentation du nombre d'échantillons n conduit à avoir des données d'apprentissage plus proches de $X = x$, la régularité de F^* est nécessaire pour garantir que l'augmentation de n est associée à une augmentation de la prédictibilité.

La définition D.1 peut être reformulée : un taux de convergence minimal optimal sur la classe $\mathcal{D}$ est un taux de convergence minimax minimal sur $\mathcal{D}$, et il existe un algorithme $\hat{F}_n$ réalisant ce taux. Nous sommes capables d'obtenir un taux de convergence minimax inférieur en utilisant une sous-classe de $\mathcal{D}^{(h,C,M)}$ qui réduit le problème à des résultats standards de régression ponctuelle (Györfi et al., 2002).

Afin de trouver un algorithme $\hat{F}_n$ qui atteigne le taux de convergence minimax minimal obtenu, nous étudions les méthodes k-nearest neighbor (k-NN) et les méthodes à noyau. La méthode k-NN est bien connue dans le cadre classique de la régression et de la classification (voir, par exemple, Biau and Devroye 2015). En régression distributionnelle, k-NN peut être adapté de manière propre pour estimer la distribution conditionnelle F_x^* et l'estimateur s'écrit comme suit

$$\hat{F}_{n,x}(z) = \frac{1}{k_n} \sum_{i=1}^{k_n} \mathbb{1}_{Y_{i:n}(x)\leq z}, \tag{D.5}$$

où $1 \leq k_n \leq n$ et $Y_{i:n}(x)$ désigne l'observation au i-ième plus proche voisin de x. Comme d'habitude, les éventuelles égalités sont rompues au hasard pour définir les plus proches voisins.

L'estimateur à noyau en régression distributionnelle (voir, par exemple, Györfi et al. 2002, Chapter 5) peut être exprimée comme suit

$$\hat{F}_{n,x}(z) = \frac{\sum_{i=1}^{n} K(\frac{x-X_i}{h_n}) \mathbb{1}_{Y_i \leq z}}{\sum_{i=1}^{n} K(\frac{x-X_i}{h_n})}, \tag{D.6}$$

si le dénominateur est différent de zéro. Lorsque le dénominateur est nul, nous utilisons la convention $\hat{F}_{n,x}(z) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{Y_i \leq z}$. La largeur de bande $h_n > 0$ dépend de la taille de l'échantillon n, et la fonction $K : \mathbb{R}^d \to [0, \infty)$ est appelée le noyau.

Nous obtenons des limites supérieures explicites et non asymptotiques pour l'excès de risque (D.2) de la méthode k-NN et de la méthode à noyau, respectivement, uniformément sur $calD^{(h,C,M)}$. L'optimisation des bornes par rapport à des choix appropriés de k_n et h_n conduit aux résultats suivants sur le taux de convergence optimal minimax.

Theorem D.1. *La suite* $a_n = n^{-\frac{2h}{2h+d}}$ *est le taux de convergence minimax optimal sur la classe* $\mathcal{D}^{(h,C,M)}$.

En particulier, les méthodes k-NN et à noyau atteignent le taux de convergence minimal optimal en dimension $d \geq 2$ et en toute dimension d, respectivement. Dans le contexte du post-traitement statistique, les méthodes k-NN et à noyau sont liées aux techniques d'analog ensemble (voir, par exemple, Delle Monache et al. 2013), et cette relation est examinée plus en détail dans l'annexe C. Des commentaires supplémentaires sur le chapitre 2 sont fournis dans l'annexe C.

Chapitre 2 : Résumé des contributions

- Nous formalisons un cadre permettant d'adapter les concepts de la théorie de l'estimation à l'étude des risques théoriques en termes de règles de score.
- Nous obtenons le taux de convergence optimal minimax en régression distributionnelle pour une classe de distributions donnée (Théorème 2.1).
- Les méthodes k-NN et à noyau atteignent le taux de convergence minimax optimal en dimension $d \geq 2$ et en toute dimension d, respectivement.
- Nous obtenons des bornes supérieures non asymptotiques sur le taux de convergence pour les méthodes k-NN et à noyau avec une taille d'échantillon fixe n (Propositions 2.1 et 2.2).
- Les résultats peuvent être étendus au CRPS pondéré par seuil (voir l'annexe C).

D.3 Distributional regression U-Nets for the postprocessing of precipitation ensemble forecasts

Opérationnellement à Météo-France, le post-traitement des prévisions de température et de précipitations repose sur des modèles locaux (c'est-à-dire un modèle par localisation) basés sur des forêts aléatoires (Taillardat and Mestre, 2020). Les forêts de régression quantiles (QRF ; Meinshausen 2006) sont une méthode non paramétrique capable de prédire des quantiles conditionnels ou, plus généralement, une distribution conditionnelle. De manière similaire aux forêts aléatoires, elles utilisent les données dans les nœuds terminaux (c'est-à-dire, les feuilles) pour calculer une moyenne pondérée des distributions empiriques. Les QRF ont prouvé leur performance pour une grande variété de variables (Taillardat et al., 2016; Whan and Schmeits, 2018; van Straaten et al., 2018; Rasp and Lerch, 2018; Taillardat et al., 2019; Schulz and

Lerch, 2022b). Les QRF sont connues pour avoir trois principales limitations : une potentielle incohérence spatiale, une voracité en mémoire de stockage (Taillardat and Mestre, 2020), et une incapacité à extrapoler.

Rasp and Lerch (2018) a proposé les réseaux de régression distributionnelle (DRN), un modèle global basé sur les réseaux de neurones (NN) qui prédit les paramètres d'une distribution d'intérêt. Il exploite la flexibilité des NN pour modéliser la dépendance des paramètres aux covariables (utilisées comme entrées du DRN). Le DRN peut être considéré comme une extension d'EMOS (Gneiting et al., 2005), qui ajuste lui-même une distribution paramétrique où les paramètres dépendent linéairement des statistiques sommaires de l'ensemble brut. Le DRN est un modèle global grâce à la présence d'un module d'embedding au sein de son architecture, permettant au réseau d'apprendre des paramètres spécifiques à une localisation et de bénéficier des données provenant de localisations similaires. Rasp and Lerch (2018) et Schulz and Lerch (2022b) ont montré que DRN surpasse les autres méthodes de pointe dans la plupart des stations en Allemagne pour le post-traitement de la température et des rafales de vent, respectivement. Bien qu'il soit un modèle global, l'architecture du DRN le rend mal adapté aux données grilles. Son architecture ne tient pas compte de la structure spatiale des points et doit donc tenter de l'apprendre via son module d'embedding.

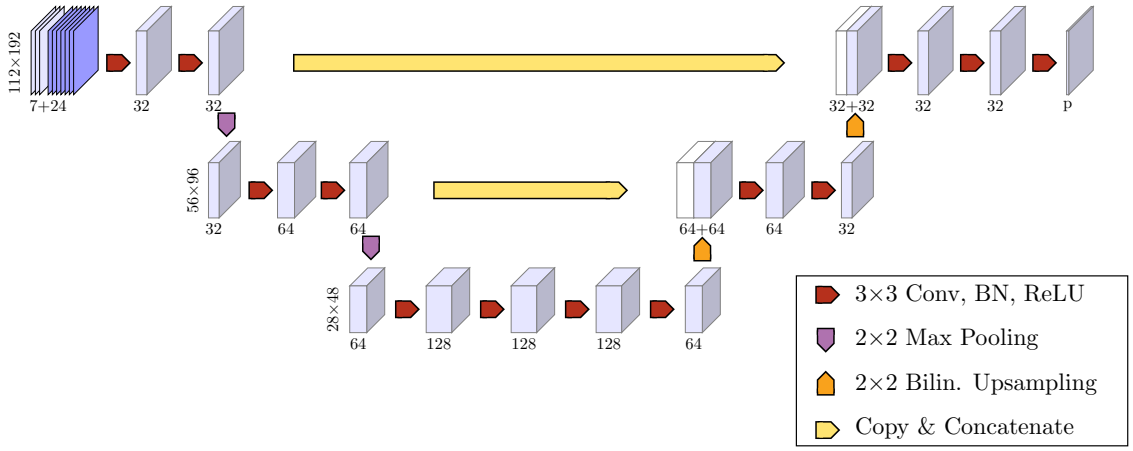


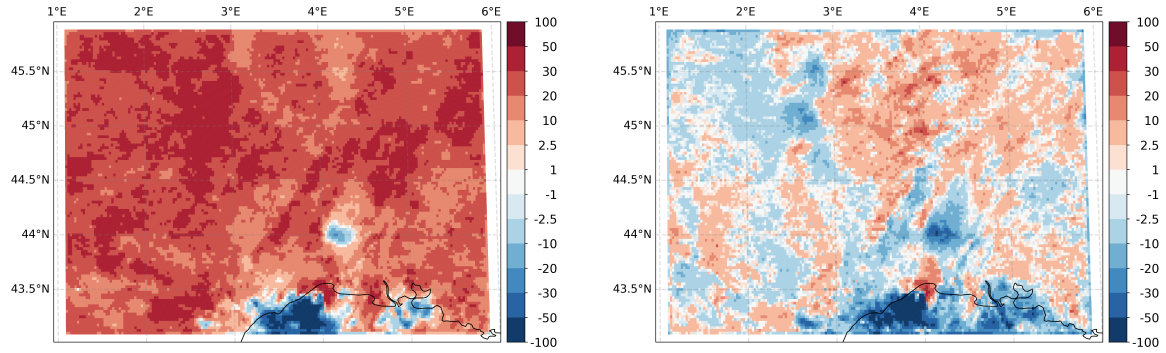
Figure D.1: Architecture des U-Nets de régression distributionnelle. *Conv* signifie convolution, *BN* signifie normalisation par lot, *ReLU* signifie rectified linear unit et *Bilin. Upsampling* signifie échantillonnage bilinéaire. p est le nombre de paramètres de distribution : pour GTCND et CSGD, $p = 3$.

Nous proposons une méthode de régression distributionnelle basée sur U-Net adaptée aux données grilles, qui prédit les paramètres des distributions marginales à chaque point de grille. La régression distributionnelle U-Net (DRU) prend en entrée à la fois des champs constants (par exemple, l'orographie) et des résumés statistiques de l'ensemble brut (voir Figure **??**). Sur la partie gauche, la succession de blocs convolutionnels spécifiques (flèches rouges et flèches violettes) conduit à une augmentation du nombre de caractéristiques et à une réduction de la dimension spatiale (c'est-à-dire un grossissement de la résolution spatiale) au fur et à mesure que les données progressent dans le réseau. Ces blocs convolutionnels sont construits afin d'apprendre des représentations utiles des caractéristiques des champs à différentes échelles spatiales. Sur la partie droite, les blocs de mise à l'échelle (flèches orange), basés sur un échantillonnage bilinéaire, utilisent les caractéristiques apprises dans la partie centrale de l'architecture pour prédire des caractéristiques à des résolutions plus fines et finalement apprendre les paramètres de la distribution sélectionnée. De plus, nous utilisons des connexions de saut (flèches jaunes)

car elles ont prouvé qu'elles améliorent la stabilité de la convergence des NN (Li et al., 2018).

Nous nous concentrons sur les précipitations accumulées sur 3 heures dans le sud de la France, qui est une région sujette aux événements de précipitations méditerranéennes intenses. Les prévisions d'ensemble sont issues du système de prévision d'ensemble à 17 membres PEAROME (Bouttier et al., 2015), qui produit un ensemble maillé sur l'Europe de l'Ouest avec une résolution horizontale de 0.025°. Les prévisions probabilistes sont comparées aux données de précipitations accumulées sur 3 heures obtenues à partir du produit radar ajusté par pluviomètre ANTILOPE (Champeaux et al., 2009). La précipitation est une grandeur météorologique difficile à prévoir en raison de sa climatologie à queue lourde et de sa dépendance spatio-temporelle à courte échelle (Hemri et al., 2014; Taillardat and Mestre, 2020). De plus, un autre aspect difficile de l'ensemble de données est que seules trois années de données d'entraînement sont disponibles. Cela est particulièrement difficile en ce qui concerne la prévision de fortes précipitations.

Pour adapter la prévision des précipitations, deux distributions paramétriques avec une masse atomique en zéro (c'est-à-dire, pour les événements secs) sont sélectionnées : la distribution normale tronquée/censurée généralisée (GTCND ; Jordan et al. 2019) et la distribution gamma décalée-censurée (CSGD ; Scheuerer and Hamill 2015a). Nous désignons U-Net+*distrib* le DRU où *distrib* est la distribution paramétrique.



(a) CRPSSde U-Net+CSGD par rapport à raw (b) CRPSS de U-Net+CSGD par rapport à QRF

Figure D.2: Performance predictive de U-Net+CSGD en terme de CRPS. CRPSS par rapport à (a) l'ensemble brut (raw) et QRF (b) de U-Net+CSGD.

Le Chapitre 3 fournit une comparaison approfondie entre l'ensemble brut, QRF, QRF avec extension de queue (TQRF ; Taillardat et al. 2019) et DRU. Ici, nous fournissons une comparaison simplifiée entre U-Net+CSGD, l'ensemble brut et QRF uniquement. En termes de CRPS, l'amélioration relative peut être exprimée en utilisant le continuous ranked probability skill score (CRPSS) défini comme

$$\mathrm{CRPSS}(F, F_{\mathrm{ref}}) = 1 - \frac{\mathbb{E}_G[\mathrm{CRPS}(F, Y)]}{\mathbb{E}_G[\mathrm{CRPS}(F_{\mathrm{ref}}, Y)]},$$

où G est la distribution des observations et $\mathbb{E}_G[\cdots]$ est l'espérance par rapport à $Y \sim G$. Le CRPSS est positif si la prévision F améliore le CRPS attendu par rapport à la prévision de référence F_{ref}, et négatif sinon. Dans ce qui suit, le CRPSS est exprimé en pourcentage. La Figure D.2 fournit le CRPSS de U-Net+CSGD par rapport à l'ensemble brut et à QRF. U-Net+CSGD conduit à un CRPSS de $22,36\%$ par rapport à l'ensemble brut en moyenne sur la région d'intérêt. Comme QRF, les DRU conduisent à une amélioration en termes de CRPSS sur la grande majorité des points de grille. Néanmoins, il existe des zones où leur performance prédictive est inférieure à celle de l'ensemble brut. Ces zones sont situées sur la mer Méditerranée ou près de la côte, et un patch est localisé dans la vallée du Rhône. Cela est dû

au fait que la zone au-dessus de la mer Méditerranée est associée aux plus faibles accumulations de précipitations et à une qualité d'observation moindre, car elle est éloignée du radar le plus proche et ne peut pas être corrigée par les pluviomètres. Globalement, U-Net+CSGD présente un CRPS attendu supérieur à celui de QRF (CRPSS moyen de $-1,37\%$), mais ils présentent une performance prédictive améliorée (en termes de CRPS) sur une partie non négligeable de la région d'intérêt. En excluant les points de grille situés au-dessus de la mer et à la frontière, le CRPSS moyen par rapport à QRF est de $0,26\%$, montrant que U-Net+CSGD a une performance prédictive comparable à QRF sur terre en termes de CRPS.

Comme mentionné dans la Section D.1.3, les histogrammes de rang sont un outil de diagnostic utile pour la calibration probabiliste des prévisions. En particulier, la planéité de l'histogramme de rang caractérise les prévisions calibrées. La planéité et d'autres formes informatives peuvent être testées statistiquement par ce que l'on appelle ici les tests Jolliffe-Primo-Zamo (JPZ) (Jolliffe and Primo, 2008; Zamo, 2016). La Figure D.3 montre les histogrammes de rang sur l'ensemble de la grille et les tests JPZ pour la planéité. Comme c'est souvent le cas, l'ensemble brut est biaisé et sous-dispersé, ce qui est visible par la forme triangulaire des histogrammes de rang et par le fait que les rangs les plus bas et les plus élevés sont surreprésentés. Le test JPZ confirme que la prévision de l'ensemble brut n'est pas calibrée (seulement 6% des points de grille ne rejettent pas la planéité de l'histogramme de rang). QRF montre une calibration très élevée avec des tests JPZ ne rejetant pas la planéité à 93% des points de grille. Les méthodes U-Net+CSGD présentent un niveau de calibration inférieur à celui de QRF, mais elles sont encore significativement calibrées. Les tests JPZ ne rejettent pas l'hypothèse de planéité pour 77% des points de grille. Les points de grille pour lesquels les prévisions U-Net+CSGD ne sont pas calibrées (c'est-à-dire les JPZ rejetant l'hypothèse de planéité) sont associés à des climatologies élevées.

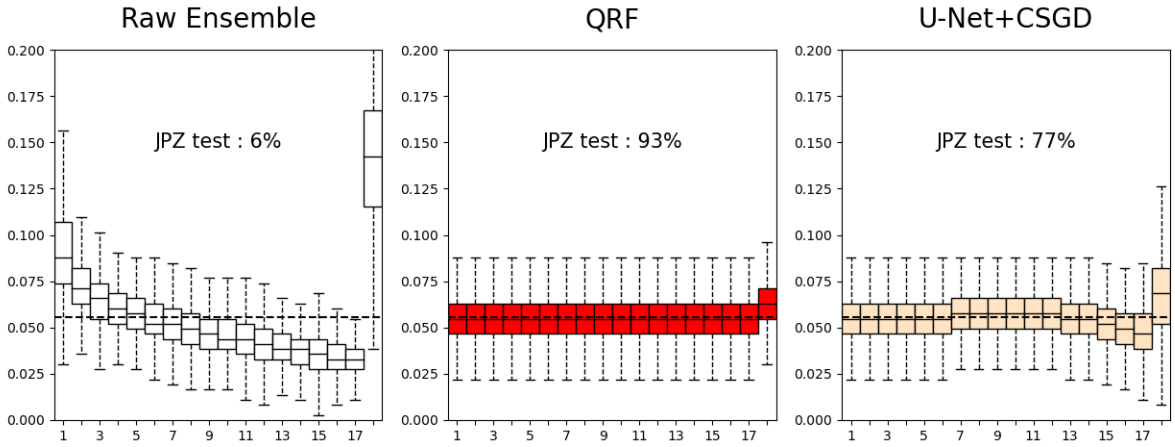


Figure D.3: Histogramme de rang pour l'ensemble brut, QRF et U-Net+CSGD.

Pour se concentrer sur la performance prédictive des prévisions concernant les événements extrêmes, nous nous intéressons à la prédiction d'événements binaires sous la forme du dépassement d'un seuil élevé t (voir Fig. D.4). Nous utilisons les courbes ROC (Receiver Operating Characteristic) pour évaluer le pouvoir discriminant des prévisions en termes de décisions binaires. En particulier, les courbes ROC peuvent informer sur le risque de manquer un événement extrême. Une bonne prévision doit maximiser le taux d'événements détectés tout en minimisant les fausses alertes. Pour des seuils élevés $t = 10$ mm et $t = 20$ mm (correspondant respectivement aux quantiles de niveau 0,995 et 0,999 de la climatologie sur la région d'intérêt), les courbes ROC des différentes méthodes de post-traitement ont un classement clair. Pour les deux seuils, la perfor-

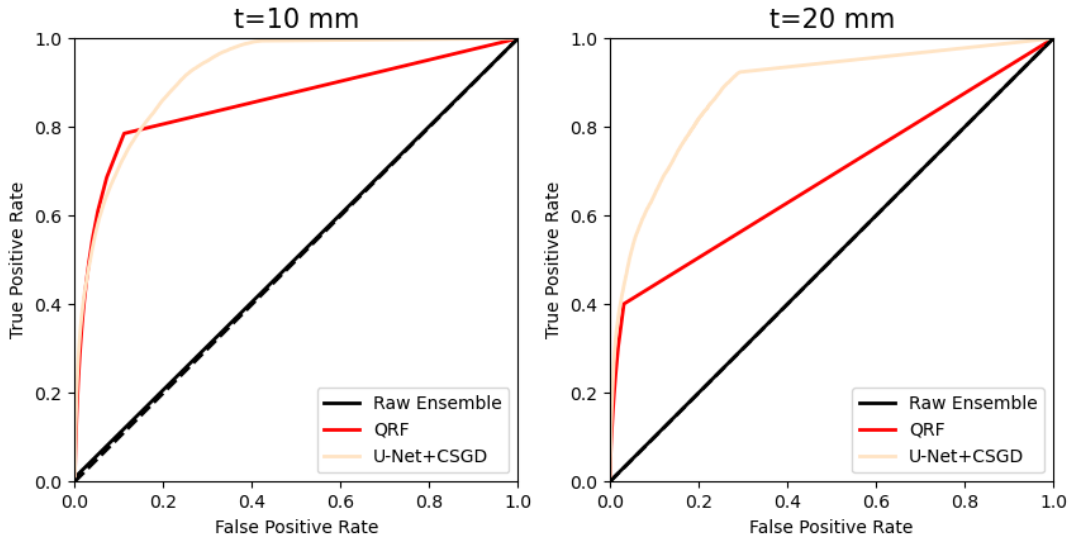


Figure D.4: Courbes *Receiver Operating Characteristic* (ROC) d'évènements binaires correspondant au dépassement d'un seuil $t = 10$ mm et $t = 20$ mm.

mance de l'ensemble brut est proche du choix aléatoire (ligne pointillée). Pour $t = 10$ mm, QRF et U-Net+CSGD sont tous deux capables de maintenir un bon pouvoir prédictif. Cependant, U-Net+CSGD a de meilleures performances que QRF. Pour $t = 20$ mm, l'écart de performance entre U-Net+CSGD et QRF continue de croître à mesure que la performance prédictive de QRF se détériore.

Nous proposons des méthodes basées sur U-Net qui peuvent simultanément post-traiter les distributions marginales à chaque point de grille en utilisant les informations des points de grille voisins. Cela permet d'éviter la voracité en mémoire de stockage et l'incapacité à extrapoler de QRF. Le DRU surpasse l'ensemble brut pour toutes les métriques utilisées. De plus, les DRU ont des performances prédictives comparables aux méthodes basées sur QRF en termes de CRPS. Les DRU sont (probabilistiquement) calibrés sur une grande partie du domaine étudié, sauf pour les zones associées à de fortes précipitations climatologiques. En ce qui concerne le pouvoir prédictif pour les fortes précipitations, U-Net+CSGD surpasse les méthodes basées sur QRF.

Chapter 3 : Résumé des contributions

- Le U-Net de régression distributionnelle (DRU) est un modèle global qui prédit des distributions paramétriques marginales et contourne certaines des limitations connues de QRF. Il fournit une extension naturelle de DRN aux données sur grille.
- Nous passons en revue les méthodes utilisant des architectures U-Net dans le post-traitement statistique (Table 3.4).
- En termes de CRPS, les performances prédictives du DRU sont comparables aux méthodes de pointe (Figure 3.5 et Table 3.5)
- Les DRU fournissent des prévisions (probabilistiquement) calibrées sur la plupart des points de grille. Cependant, ils échouent dans les zones de fortes précipitations climatologiques (Figures 3.8 et 3.9).
- U-Net+CSGD surpasse les autres méthodes en termes de dépassement des seuils de précipitations élevées (Figure 3.10).

D.4 Proper scoring rules for multivariate probabilistic forecasts based on aggregation and transformation

La section précédente (et le chapitre associé) ne post-traite pas la structure de dépendance de la quantité d'intérêt, supposant qu'elle peut être récupérée à partir de l'ensemble brut (par exemple, en utilisant ECC ; Schefzik et al. 2013) ou à partir de la climatologie (par exemple, en utilisant ScS ; Clark et al. 2004) ou qu'elle peut être traitée séparément comme mentionné dans la Section D.1. Néanmoins, c'est un aspect crucial des prévisions, car cela peut influencer l'impact d'un événement. Les prévisions probabilistes spatiales nécessitent des méthodes de vérification propres.

Les règles de score sont un outil de choix pour quantifier à quel point une prévision est bonne et pour comparer des prévisions concurrentes. Rappelons que la propreté permet d'évaluer simultanément la calibration et la sharpness (Winkler, 1977; Winkler et al., 1996); ainsi, elle encourage les prévisionnistes à suivre leurs véritables croyances et prévient les contournements. Cependant, c'est une propriété nécessaire des bonnes règles de score, mais cela ne garantit pas qu'une règle de score fournit une caractérisation informative de la performance prédictive. En particulier, la propreté ne garantit pas que les prévisions minimisant le score attendu sont pertinentes pour la tâche à accomplir. Même la stricte propriété ne garantit pas que les prévisions à proximité du score attendu minimum soient proches de la prévision idéale. Dans les contextes univariés et multivariés, de nombreuses études ont prouvé qu'aucune règle de score n'a tout, et donc, différentes règles de score devraient être utilisées pour mieux comprendre la performance prédictive des prévisions (voir, par exemple, Scheuerer and Hamill 2015b; Taillardat 2021; Bjerregård et al. 2021).

Cela peut expliquer le développement des *méthodes de vérification spatiale* (Gilleland et al., 2009; Dorninger et al., 2018), qui sont des méthodes de vérification basées sur la physique pour les prévisions spatiales. Elles reposent sur la robustesse face à l'effet de double pénalité (Ebert, 2008) et sur l'interprétabilité à la fois des valeurs uniques et du classement des prévisions. Cependant, la grande majorité des méthodes ne sont pas propres. Dans le contexte des règles de score propres, l'interprétabilité peut découler d'une fonction de score consistante pour une fonctionnelle (par exemple, l'erreur quadratique est induite par une fonction de score consistante pour la moyenne ; Gneiting 2011), de la connaissance des aspects de la prévision que la règle de score discrimine (par exemple, la règle de score de Dawid-Sebastiani ne discrimine les prévisions que par leur moyenne et leur variance ; Dawid and Sebastiani 1999) ou de la connaissance des limites d'une certaine règle de score propre (par exemple, le variogram score est incapable de discriminer deux prévisions qui ne diffèrent que par un biais constant ; Scheuerer and Hamill 2015b). Dans ce contexte, les règles de score propres et interprétables deviennent des méthodes de vérification de choix car le classement des prévisions qu'elles produisent peut être plus informatif que le classement d'une règle de score plus complexe mais moins interprétable.

Scheuerer and Hamill (2015b) a proposé le variogram score pour cibler la vérification de la structure de dépendance. Le variogram score d'ordre p ($p > 0$) est définie comme

$$\mathrm{VS}_p(F, \boldsymbol{y}) = \sum_{i,j=1}^{d} w_{ij} \left(\mathbb{E}_F \left[|X_i - X_j|^p \right] - |y_i - y_j|^p \right)^2 ,$$

où X_i est le i-ème composant du vecteur aléatoire $\boldsymbol{X} \in \mathbb{R}^d$ suivant F, les w_{ij} sont des poids non négatifs, et $\boldsymbol{y} \in \mathbb{R}^d$ est une observation. La construction du variogram score repose sur deux principes. Tout d'abord, le variogram score est la somme pondérée de règles de score agissant sur la distribution de $\boldsymbol{X}_{i,j} = (X_i, X_j)$ et sur des composants appariés des observations $y_{i,j}$. Ce principe d'*agrégation* permet la combinaison de règles de score propres et les résume en une règle de score propre agissant sur l'ensemble de la distribution F et des observations

$\boldsymbol{y}$. Deuxièmement, les règles de score composant la somme pondérée peuvent être considérées comme une règle de score propre standard appliquée à des transformations des prévisions et des observations. Soit $\gamma_{i,j} : \boldsymbol{x} \mapsto |x_i - x_j|^p$ la transformation liée au variogramme d'ordre p, alors le variogram score peut être réécrit comme

$$\mathrm{VS}_p(F, \boldsymbol{y}) = \sum_{i,j=1}^{d} w_{ij} \mathrm{SE}(\gamma_{i,j}(F), \gamma_{i,j}(\boldsymbol{y})),$$

où $\mathrm{SE}(F, y) = (\mathbb{E}_F[X] - y)^2$ est l'erreur quadratique univariée (SE) et $\gamma_{i,j}(F)$ est la distribution de $\gamma_{i,j}(\boldsymbol{X})$ pour $\boldsymbol{X} \sim F$. Ce deuxième principe est le principe de *transformation*, permettant de construire des règles de score propres basées sur des transformations qui peuvent bénéficier de l'interprétabilité découlant d'une transformation (ici, la transformation de variogramme $\gamma_{i,j}$) et de la simplicité et de l'interopérabilité des règles de score propres sur lequelles elles s'appuient (ici, la SE).

Ces deux principes ont été diffusés dans la littérature au cours des dernières décennies. Plus explicitement, Dawid and Musio (2014) propose la notion de *score composite*, qui est un cas particulier de la combinaison des deux principes. Heinrich-Mertsching et al. (2024) introduit le principe de transformation et l'applique dans le contexte des processus ponctuels. Nous formalisons des formes générales des principes d'agrégation et de transformation et leur combinaison conduit au Corollaire D.1.

Corollary D.1. *Soit $\mathcal{T} = \{T_i\}_{1 \leq i \leq m}$ un ensemble de transformations de $\mathbb{R}^d$ vers $\mathbb{R}^k$. Soit $\mathcal{S}_\mathcal{T} = \{\mathrm{S}_{T_i}\}_{1 \leq i \leq m}$ un ensemble de règles de score propres où S est propre par rapport à $T_i(\mathcal{F})$, pour tout $1 \leq i \leq m$. Soit $\boldsymbol{w} = \{w_i\}_{1 \leq i \leq m}$ des poids non négatifs. Alors, la règle de score*

$$\mathrm{S}_{\mathcal{S}_\mathcal{T}, \boldsymbol{w}}(F, \boldsymbol{y}) = \sum_{i=1}^{m} w_i \mathrm{S}_{T_i}(F, \boldsymbol{y})$$

est propre par rapport à $\mathcal{F}$.

Pour gagner en interprétabilité, il est naturel d'avoir des transformations réduisant la dimension (c'est-à-dire, $k < d$) car cela conduit à des transformations simplifiant les quantités multivariées. En particulier, il est généralement préférable de choisir $k = 1$ pour rendre la quantité plus facile à interpréter et se concentrer sur des informations spécifiques contenues dans la prévision ou l'observation. De plus, nous montrons que toutes les règles de score de noyau peuvent être exprimées comme l'agrégation de SE appliquées à une séquence de transformations.

Les règles de score basées sur l'agrégation et la transformation peuvent tirer parti de l'interprétabilité à la fois des transformations et des règles de score standard. Par exemple, si l'intérêt porte sur la performance prédictive des prévisions en termes de leur prédiction du dépassement d'un seuil t, le Brier score (Brier, 1950) devrait être utilisé dans un cadre univarié. Le Brier score s'exprime comme

$$\mathrm{BS}_t(F, y) = ((1 - F(t)) - \mathbb{1}_{y>t})^2 = (F(t) - \mathbb{1}_{y \leq t})^2,$$

où $1 - F(t)$ est la probabilité prédite que le seuil t soit dépassé. L'espérance du Brier score est minimale pour toutes les prévisions F telles que la probabilité de dépassement du seuil t est correctement prédite. Dans un contexte de vérification spatiale, le dépassement du seuil à chaque emplacement peut être résumé par le Brier score agrégé

$$\frac{1}{d} \sum_{i=1}^{d} \mathrm{BS}(F_i, y_i),$$

où F_i est la distribution marginale de F à l'emplacement i et y_i est la valeur de $\boldsymbol{Y}$ à l'emplacement i. Dans ce cas, les transformations sont des projections sur chaque emplacement (c'est-à-dire, les marges unidimensionnelles) et l'agrégation utilise des poids uniformes puisque aucune hypothèse n'est faite sur les emplacements.

Lorsqu'on considère la structure de dépendance spatiale, une quantité d'intérêt peut être le dépassement d'un seuil t à des emplacements voisins. Dans le cas des précipitations, les voisinages pourraient être définis comme des bassins versants de rivières. La fraction de dépassement de seuil (FTE) est la statistique résumée associée au dépassement simultané d'un certain seuil et elle est définie comme

$$\mathrm{FTE}_{P,t}(\boldsymbol{X}) = \frac{1}{|P|}\sum_{i\in P}\mathbb{1}_{\{X_i\geq t\}},$$

où P est un patch (ou un voisinage) d'intérêt et $|P|$ sa dimension. En utilisant les principes d'agrégation et de transformation, la SE agrégée de la FTE est défini comme

$$\sum_{P\in\mathcal{P}} w_P \mathrm{SE}\big(\mathrm{FTE}_{P,t}(F), \mathrm{FTE}_{P,t}(\boldsymbol{y})\big) = \sum_{P\in\mathcal{P}} w_P \big(\mathbb{E}_F[\mathrm{FTE}_{P,t}(X)] - \mathrm{FTE}_{P,t}(\boldsymbol{y})\big)^2$$

où $\mathcal{P}$ est un ensemble de patchs, w_P est le poids associé à un patch $P \in \mathcal{P}$. Cette règle de score est propre et se concentre sur la prédiction du dépassement d'un seuil t via la fraction de lieux dans un patch P dépassant ce seuil. La ressemblance avec le Brier score est claire et l'SE agrégé de la FTE devient le BS agrégé lorsque des patchs contenant un seul emplacement sont considérées.

De nombreux autres exemples de transformations (et des règles de score qui en résultent) sont présentés, discutés et liés à la littérature dans le Chapitre 4. Plusieurs expériences numériques sont développées pour illustrer l'importance de l'interprétabilité dans un cadre pratique et, plus particulièrement, comment les règles de score basées sur l'agrégation et la transformation peuvent être utilisées dans la vérification des prévisions spatiales. En particulier, nous montrons comment les règles de score habituelles peuvent être adaptées pour éviter l'effet de double pénalité.

Chapitre 4 : Résumé des contributions

- Nous fournissons une revue complète des règles de score univariées et multivariées à la lumière de l'interprétabilité (Section 4.2).
- Nous formalisons un cadre théorique (présent dans la littérature) basé sur les principes d'agrégation et de transformation pour construire des règles de score propres interprétables (Section 4.3).
- Les scores de noyau peuvent être exprimés comme une agrégation d'erreurs quadratiques appliquées à une suite de transformations (Proposition 4.3).
- Nous listons des exemples de règles de score basées sur l'agrégation et la transformation provenant à la fois de la littérature et de suggestions originales (Section 4.4).
- Des expériences numériques ont été réalisées pour illustrer les avantages des règles de score propres interprétables dans divers contextes (Section 4.5).
- En particulier, des solutions concrètes sont données pour aider à combler le fossé avec les outils de vérification spatiale.

D.5 Travaux associés

L'Appendice A et l'Appendice B reproduisent deux travaux connexes réalisés durant la thèse (Dombry et al., 2024; Al Masry et al., 2023). Ces travaux sont liés à la régression distributionnelle mais ne sont pas directement connectés au post-traitement statistique. Ci-après, nous motivons brièvement ces travaux, explicitons leur relation avec les travaux présentés dans les sections précédentes et résumons leurs principales contributions.

D.5.1 Stone's theorem for distributional regression in Wasserstein distance

Comme mentionné précédemment, le Chapitre 2 adapte des résultats de la régression ponctuelle à la régression distributionnelle. Au lieu de viser une convergence optimale pour une classe donnée de distributions, l'Appendice A se concentre sur la consistance universelle en régression distributionnelle (c'est-à-dire, des résultats de convergence valables pour toute distribution mais sans garantie sur le taux de convergence).

Rappelons le cadre général de la régression introduit dans la Section D.2. Nous observons un échantillon $D_n = \{(X_i, Y_i), 1 \leq i \leq n\}$, de copies indépendantes de $(X, Y) \in \mathbb{R}^d \times \mathbb{R}^m$ avec distribution P. Sur la base de cet échantillon et en supposant que Y est intégrable, en régression ponctuelle, l'objectif est d'estimer la fonction de régression

$$r(x) = \mathbb{E}[Y|X = x], \quad x \in \mathbb{R}^d.$$

Les estimateurs de moyenne locale prennent la forme

$$\hat{r}_n(x) = \sum_{i=1}^{n} W_{ni}(x) Y_i$$

avec $W_{n1}(x), \ldots, W_{nn}(x)$ les *poids locaux* à x. Les poids locaux sont supposés être des fonctions mesurables de x et de $X_1, \ldots, X_n$ mais ne pas dépendre de $Y_1, \ldots, Y_n$. Considérons le cas de poids de probabilité satisfaisants

$$W_{ni}(x) \geq 0, \ 1 \leq i \leq n, \quad \text{et} \quad \sum_{i=1}^{n} W_{ni}(x) = 1. \tag{D.7}$$

Le théorème de Stone stipule la consistance universelle de l'estimation de régression dans la norme L^p.

Theorem D.2 (Stone 1977)**.** *Supposons que les poids de probabilité* (D.7) *satisfassent les trois conditions suivantes :*

i) il existe $C > 0$ *tel que* $\mathbb{E}\big[\sum_{i=1}^{n} W_{ni}(X) g(X_i)\big] \leq C\mathbb{E}[g(X)]$ *pour tout* $n \geq 1$ *et toute fonction mesurable* $g : \mathbb{R}^k \to [0, +\infty)$ *telle que* $\mathbb{E}[g(X)] < \infty$ *;*

ii) pour tout $\varepsilon > 0$, $\sum_{i=1}^{n} W_{ni}(X) \mathbb{1}_{\{\|X_i - X\| > \varepsilon\}} \to 0$ *en probabilité lorsque* $n \to +\infty$ *;*

iii) $\max_{1 \leq i \leq n} W_{ni}(X) \to 0$ *en probabilité lorsque* $n \to +\infty$*.*

Alors, pour tout $p \geq 1$ *et* $(X, Y) \sim P$ *tel que* $\mathbb{E}[\|Y\|^p] < \infty$*,*

$$\mathbb{E}\big[\|\hat{r}_n(X) - r(X)\|^p\big] \longrightarrow 0 \quad \textit{quand } n \to +\infty. \tag{D.8}$$

Inversement, si (D.8) *est vraie, alors les poids de probabilité doivent satisfaire les conditions* $i) - iii)$*.*

Des exemples d'estimateurs de moyenne locale incluent les k-NN, les méthodes à noyau et certaines variantes des forêts aléatoires.

Pour adapter ce résultat à la régression distributionnelle, nous utilisons une définition de convergence basée sur la distance de Wasserstein plutôt que sur les règles de score, contrairement au Chapitre 2 :

$$\mathbb{E}\left[\mathcal{W}_p^p(\hat{F}_{n,X}, F_X^*)\right] \longrightarrow 0 \quad \text{quand } n \to +\infty, \tag{D.9}$$

où $\mathcal{W}_p$ est la distance de Wasserstein d'ordre p et $\hat{F}_{n,X}$ est une estimation de la distribution conditionnelle de Y donné X, notée F_X^*. Considérons l'estimateur de distribution empirique pondérée basé sur l'échantillon d'entraînement D_n

$$\hat{F}_{n,X} = \sum_{i=1}^{n} W_{ni}(X)\delta_{Y_i}, \tag{D.10}$$

où δ_y désigne la masse de Dirac au point $y \in \mathbb{R}^m$.

En utilisant la notion de max-sliced distance de Wasserstein (Bayraktar and Guo, 2021), ce travail étend le théorème de Stone à la régression distributionnelle dans la distance de Wasserstein d'ordre $p \geq 1$. Plus précisément, il existe une équivalence entre le fait que les poids de l'estimateur de distribution empirique pondérée (D.10) satisfont les conditions $i) - iii)$ et la convergence de (D.9). De plus, pour $p = 1$, les taux de convergence minimax optimaux sur des classes spécifiques de distributions sont obtenus. Les applications du théorème de Stone en régression distributionnelle sont illustrées à l'aide de l'estimation de l'espérance conditionnelle de la queue et des moments pondérés par probabilité, entre autres.

Appendix A : Résumé des contributions

- Nous adaptons le théorème de Stone à la régression distributionnelle (multidimensionnelle) : consistance universelle en termes de distance de Wasserstein d'ordre $p \geq 1$ dans un cadre multivarié (Théorème A.2).
- Nous déterminons les taux de convergence minimax optimaux pour $p = 1$ et $m = 1$ (Théorème A.3).
- Nous illustrons les applications du théorème de Stone aux statistiques récapitulatives en régression distributionnelle (Section A.3.3).

D.5.2 A new methodology to predict the oncotype scores based on clinicopathological data with similar tumor profiles

L'Appendice B utilise une technique de régression distributionnelle pour aider les cliniciens dans leur prise de décision concernant la prédiction du risque de récidive du cancer du sein et des traitements potentiels.

Le test Oncotype DX (ODX) est un test moléculaire commercialement disponible pour l'analyse du cancer du sein qui fournit des informations pronostiques et prédictives sur le risque de récidive du cancer du sein pour les patientes hormonodépendantes et HER2-négatif. Le test ODX fournit un score de récidive (score ODX) compris entre 0 et 100. Des valeurs plus élevées du score ODX correspondent à un risque de récidive plus élevé. Plusieurs études rétrospectives et prospectives ont validé ce test et son utilité clinique (voir, par exemple, Paik et al. 2004,?, 2006; Albain et al. 2010).

Les interprétations les plus courantes du score ODX passent par des seuils définissant deux ou trois classes de risque : par exemple, faible risque < 11, risque intermédiaire 11-25 et risque

élevé > 25 (Sparano et al., 2018). Malgré son utilité clinique, le test ODX est coûteux et le score ODX manque d'explicabilité. Afin de contourner ces limitations, des études ont tenté d'utiliser des caractéristiques clinico-pathologiques pour prédire le score ODX, soit par sa valeur directe, soit par une classification en termes de niveaux de risque. De nombreuses méthodes d'apprentissage statistique ont été étudiées, telles que la régression linéaire multiple (Klein et al., 2013; Hou et al., 2017), les forêts aléatoires (Kim et al., 2019; Pawloski et al., 2021) et les réseaux de neurones (Kim et al., 2019; Baltres et al., 2020).

Afin d'avoir une connaissance complète de l'incertitude, nous avons proposé de prédire la distribution complète de l'ODX conditionnellement aux caractéristiques clinico-pathologiques : la forêt de régression distributionnelle (DRF ; Meinshausen 2006; Athey et al. 2019). Étant donné que la DRF fournit une prédiction probabiliste, sa sortie peut prendre la forme d'une fonction de densité de probabilité discrète. De plus, elle peut être résumée par des quantités plus compréhensibles pour les praticiens, telles que sa moyenne et son écart-type ou les probabilités des classes d'intérêt, tirant parti de la familiarité des praticiens avec les interprétations du score ODX. En particulier, les patients ayant des profils similaires (en termes de poids de la forêt) peuvent être utilisés pour informer les praticiens des patientes similaires présentes dans la cohorte et détecter les prédictions non informatives liées à un manque de représentativité. Par ailleurs, la DRF a une performance comparable aux méthodes précédemment proposées dans la littérature.

Appendix B : Résumé des contributions

- Nous utilisons une technique de régression distributionnelle (DRF) pour prédire le risque de récidive du cancer du sein et fournir des informations utiles à la prise de décision.
- Les DRF fournissent une prédiction probabiliste qui peut être résumée en quantités compréhensibles pour les praticiens (par exemple, un voisinage de patients proches dans la cohorte) (Figure B.2).
- En termes de classification à faible risque et à haut risque, les DRF atteignent une performance prédictive comparable aux méthodes de pointe (Tableau B.3).

ECOLE DOCTORALE
CARNOT - PASTEUR

Titre : Post-traitement statistique des prévisions d'ensemble : théorie, application en météorologie et vérification
Mots clés : prévisions probabilistes, machine learning, régression distributionelle, scoring rules

Résumé : Cette thèse porte sur l'utilisation de méthodes de post-traitement statistiques dans le but d'améliorer les prévisions d'ensemble. Les prévisions d'ensemble sont des prévisions composées de différents membres dont la diversité tente de capturer l'incertitude liée à la prédiction. Les prévisions d'ensemble souffrent de biais et de sous-dispersion et un post-traitement est donc nécessaire afin d'améliorer leur performance. D'un point de vue théorique, cette thèse apporte des résultats sur le taux de convergence en régression distributionnelle en termes de *continuous ranked probability score*. De nombreuses méthodes de post-traitement statistiques établies tombent dans le cadre théorique de ce résultat. Par ailleurs, dans le cadre de la collaboration avec Météo-France, une méthode de post-traitement statistique basée sur un réseau de neurones U-Net a été développée afin de remédier aux limitations des méthodes utilisées opérationnellement lors de l'utilisation de données sous forme de grille. Cette thèse propose également la construction de *scoring rules* propres basées sur l'agrégation et la transformation afin de faciliter la vérification de prévisions probabilistes dans le cas multivarié.
En parallèle de la thématique principale de cette thèse, des travaux portant sur la convergence universelle en régression distributionnelle et sur l'utilisation de méthodes de régression distributionnelle pour prédire le risque de récurrence de cancer du sein ont été conduits.

Title : Statistical postprocessing of ensemble forecasts: theory, application in weather forecasting and verification
Keywords : probabilistic forecast, machine learning, distributional regression, scoring rules

Abstract : This thesis focuses on the use of statistical postprocessing methods to improve ensemble forecasts. Ensemble forecasts are composed of different members whose diversity attempts to capture forecast uncertainties. Ensemble forecasts suffer from bias and underdispersion, and postprocessing is therefore necessary to improve their performance. From a theoretical point of view, this thesis provides rates of convergence for distributional regression in terms of continuous ranked probability score. Numerous well-established statistical post-processing methods fall within the theoretical framework of this result. Furthermore, in collaboration with Météo-France, a statistical postprocessing method based on a U-Net neural network has been developed to overcome the limitations of the methods used operationally when using gridded data. This thesis also proposes the construction of proper scoring rules based on aggregation and transformation to facilitate the verification of probabilistic forecasts in multivariate settings.
In parallel with the main theme of this thesis, work has been carried out on universal convergence in distributional regression and on the use of distributional regression methods to predict breast cancer recurrence risk.

Codes AMS : 62C05, 62F99, 62G08, 62H11, 62P12, 62P10, 62-04.

UBFC UNIVERSITÉ BOURGOGNE FRANCHE-COMTÉ

Université Bourgogne Franche-Comté
32, avenue de l'Observatoire
25000 Besançon