Artificial Intelligence Review (2024) 57:94
https://doi.org/10.1007/s10462-023-10698-8

A review of predictive uncertainty estimation with machine learning

Hristos Tyralis[1,2] · **Georgia Papacharalampous**[1]

Accepted: 31 December 2023 / Published online: 18 March 2024

Abstract
Predictions and forecasts of machine learning models should take the form of probability distributions, aiming to increase the quantity of information communicated to end users. Although applications of probabilistic prediction and forecasting with machine learning models in academia and industry are becoming more frequent, related concepts and methods have not been formalized and structured under a holistic view of the entire field. Here, we review the topic of predictive uncertainty estimation with machine learning algorithms, as well as the related metrics (consistent scoring functions and proper scoring rules) for assessing probabilistic predictions. The review covers a time period spanning from the introduction of early statistical (linear regression and time series models, based on Bayesian statistics or quantile regression) to recent machine learning algorithms (including generalized additive models for location, scale and shape, random forests, boosting and deep learning algorithms) that are more flexible by nature. The review of the progress in the field, expedites our understanding on how to develop new algorithms tailored to users' needs, since the latest advancements are based on some fundamental concepts applied to more complex algorithms. We conclude by classifying the material and discussing challenges that are becoming a hot topic of research.

Keywords Boosting · Deep learning · Distributional regression · Ensemble learning · Machine learning · Probabilistic forecasting · Quantile regression · Random forests

1 Introduction

In the vast majority of supervised machine learning applications, point predictions are made that are intended to be close to the actual values of continuous processes. Point predictions are single values that represent the best estimate of a future outcome, based on

✉ Hristos Tyralis
hristos@itia.ntua.gr; montchrister@gmail.com

Georgia Papacharalampous
papacharalampous.georgia@gmail.com

[1] Department of Topography, School of Rural, Surveying and Geoinformatics Engineering, National Technical University of Athens, Iroon Polytechniou 5, 157 80 Zografou, Greece

[2] Construction Agency, Hellenic Air Force, Mesogion Avenue 227–231, 15 561 Cholargos, Greece

a set of historical data. They are often used in supervised machine learning applications where the aim is to predict a continuous variable, such as the temperature in weather forecasting or the next day's stock price. Machine learning regression algorithms can optimize a squared error (or similar) loss function to issue point predictions. Although point predictions are useful, the information content can be increased when predictions take the form of probability distributions (Dawid 1984; Gneiting and Raftery 2007). In this case, the predictive uncertainty is estimated, and better informed decisions under uncertainty can be made.

Notwithstanding that statistical modelling is mostly associated with inference of parameters, it has been indicated that the ultimate goal should be the prediction of future events (Billheimer 2019). The earliest formal strategy for estimating the probability distribution of predictions consists of fitting a Bayesian statistical model to given data (Roberts 1965; Barbieri 2015). The Bayesian problem is formulated by assigning a prior distribution to the parameters of the statistical model, updating the distribution of the parameters conditional on the data and estimating the predictive uncertainty by integrating the parameters posterior distribution. The probability distribution of the predictions is called predictive distribution, while the procedure can be called predictive uncertainty estimation (or predictive uncertainty quantification), since probability distributions characterize the uncertainty of the predictions. Although earlier considerations on predictive uncertainty estimation were Bayesian statistical-based for independent and identically distributed (IID) variables, further advancements were made possible due to the relevant progress in time series forecasting, mostly again in Bayesian settings (Chatfield 1996).

In a distinct direction, advancements became also possible due to progress in decision theory through the advent of new loss functions following the theory of consistent scoring functions and proper scoring rules (Gneiting and Raftery 2007; Gneiting 2011a; Dawid and Musio 2014). An indicative example is the quantile loss function (Koenker and Bassett Jr 1978), which can be used to estimate quantiles of the predictive distribution, when it is minimized in a regression setting.

The machine learning paradigm is grounded on the cross-validation technique, in which the parameters of the model are estimated with respect to minimizing a loss function. There is a consensus that machine learning models are more accurate compared to simple statistical models, when predicting responses due to their flexible nature (Breiman 2001b; Shmueli 2010). Naturally, the advent of new loss functions tailored to estimate probability distributions, combined with the progress in the field of machine learning algorithms, lead to machine learning algorithms that can estimate predictive uncertainty and can be more accurate compared to simpler statistical models.

Here, we review the topic of probabilistic forecasting and prediction using machine learning algorithms. "Prediction" is a general term in the sense that it encompasses the term "forecasting". The latter term refers to the case that one predicts the future conditional on past events, and is regularly accompanied by the concept of temporal dependence. Machine learning algorithms developed to model IID variables can also be applied to forecasting problems in a straightforward way and frequently show good performance, despite not exploiting temporal dependence information. We decided to distinguish the fields of probabilistic prediction and forecasting here, because numerous significant developments have originated in the latter field. Of course, ideas from one field can be transferred directly to the other field; therefore they are introduced in a unifying way hereinafter.

We aim to give an overview of the main concepts, methodologies and research techniques for predictive uncertainty estimation with machine learning algorithms.

Moreover, we aim to provide insight on how seemingly unrelated features can be combined to form new algorithms for probabilistic prediction tailored to users' needs, by synthesizing the literature. To this end, we review the following components of probabilistic prediction algorithms:

(a) Concepts related to metrics (consistent scoring functions and proper scoring rules) for assessing probabilistic predictions.
(b) Bayesian settings.
(c) Simpler as well as more complex statistical and machine learning models.

Those components can form new machine learning algorithms, when they are combined. Furthermore, we examine some special cases of probabilistic prediction including combinations of techniques, time series forecasting, spatial prediction, prediction of extremes and measurement errors. These special cases appear less frequently in the literature (with the exception of combinations of techniques and time series forecasting) and seem to be a subject of timely debate. Challenges in probabilistic prediction are also discussed.

The remainder of the manuscript is structured as follows. Section 2 presents definitions related to probabilistic forecasting and prediction that are associated with Bayesian theory, as well as the definition of the problem of probabilistic prediction in regression settings. Moreover, we list past review papers that surveyed parts of our research topic, and specify how this review paper advances the field and how it differs from the existing literature. The theory of metrics for assessing probabilistic predictions follows in Sect. 3, while Sects. 4 and 5 present statistical and machine learning algorithms for probabilistic prediction, respectively. The metrics section comes before the algorithms section because algorithms are constructed based on loss functions. Section 6 is dedicated to neural networks and deep learning for probabilistic prediction, due to the increasing number of relevant applications using those algorithms. To understand the concepts presented in the aforementioned sections, we demonstrate an example of probabilistic prediction using simulated data in Sect. 7. Ensemble learning, also termed as combination of predictions in the statistical literature, has been proven to increase predictive performance compared to base learners. We present ensemble learning concepts for probabilistic predictions in Sect. 8. Special applications of probabilistic predictions that are of interest to the wider scientific community, including among others, temporal, spatial and spatiotemporal prediction, prediction of extreme events, and uncertainty in measurements are presented in Sect. 9, while related applications in various scientific fields are listed in Sect. 9.6. The manuscript concludes with a synthesis of results and future outlook in Sect. 10.

2 Definitions and history

2.1 Bayesian statistical modelling

Let $\underline{y}$ be the response (dependent) variable of a regression model and let $\underline{x}$ be the p-dimensional vector of predictor (explanatory) variables. Note also that $\underline{y}$ and $\underline{x}$ are random variables, while we underlie them to distinguish them from respective observations. A regression model that expresses the relationship between $\underline{x}$ and $\underline{y}$ can be defined by the distribution $F_{\underline{y}|\boldsymbol{\theta},\mathbf{x}}$ of the random variable $\underline{y}$ given a parameter vector $\boldsymbol{\theta}$ and the realization $\underline{\mathbf{x}}$:

$$F_{\underline{y}|\boldsymbol{\theta},\boldsymbol{x}}(y|\boldsymbol{\theta},\boldsymbol{x}) := P\left(\underline{y} \leq y|\boldsymbol{\theta},\boldsymbol{x}\right) \tag{1}$$

The prediction problem is then defined as follows: Given the form of the statistical model F, realizations $(y|\mathbf{x},\boldsymbol{\theta})$ of the vector $(\underline{y}, \underline{x})$ and future yet unobserved realizations $\widetilde{\underline{x}}$ of $\underline{x}$, estimate the distribution $p(z|y, \boldsymbol{x}, \widetilde{\underline{x}})$ of future predictions $\underline{z}$ of $\underline{y}$. The distribution $p(z|y, \boldsymbol{x}, \widetilde{\boldsymbol{x}})$ is called predictive distribution (Gelman et al. 2013; Barbieri 2015) and a review of methods for its estimation with machine learning algorithms is the topic of our article.

If $\boldsymbol{\theta}$ was known, then it would be straightforward to estimate the predictive distribution using Eq. (1). Unfortunately, in practical situations, that is not the case, consequently the parameter vector $\boldsymbol{\theta}$ has to be estimated from the data. However, a point estimate of $\boldsymbol{\theta}$ (e.g., a maximum likelihood estimate) would ignore the estimation uncertainty, therefore, it seems reasonable to treat $\boldsymbol{\theta}$ as a random variable. Here, Bayesian statistical modelling comes into play. Relevant material regarding predictive inference can be found in some early works in Bayesian statistics (see, e.g., the expository study on univariate distributions by Roberts 1965, as well as the summary by Barbieri 2015), while Bayesian theory can be found in a large list of Bayesian books (Robert 2007; Bernardo and Smith 2008; Gelman et al. 2013).

We return to our regression problem by defining the Bayesian statistical regression model made by the statistical model $F_{\underline{y}|\boldsymbol{\theta},\boldsymbol{x}}(y|\boldsymbol{\theta}, \boldsymbol{x})$ with respective probability density function $f_{\underline{y}|\boldsymbol{\theta},\mathbf{x}}(y|\boldsymbol{\theta}, \mathbf{x})$ and a prior distribution on the parameters $p(\boldsymbol{\theta})$. Now given realizations $(y, \boldsymbol{x})$ of the vector $(\underline{y}, \underline{\mathbf{x}})$, it is possible to estimate $p(\boldsymbol{\theta}|y, \boldsymbol{x})$, called posterior parameter distribution:

$$p(\boldsymbol{\theta}|y,\boldsymbol{x}) = p(\boldsymbol{\theta}) f_{\underline{y}|\boldsymbol{\theta},\boldsymbol{x}}(y|\boldsymbol{\theta},\boldsymbol{x}) / \int p(\boldsymbol{\theta}) f_{\underline{y}|\boldsymbol{\theta},\boldsymbol{x}}(y|\boldsymbol{\theta},\boldsymbol{x}) \mathrm{d}\boldsymbol{\theta} \tag{2}$$

Given realizations $(y, \boldsymbol{x})$ and future realizations $\widetilde{\boldsymbol{x}}$ of $\boldsymbol{x}$, the predictive distribution of the response variable $\underline{z}$ is:

$$p(z|y,\boldsymbol{x},\widetilde{\boldsymbol{x}}) = \int f_{\underline{y}|\boldsymbol{\theta},\boldsymbol{x}}\left(z|\boldsymbol{\theta},\widetilde{\boldsymbol{x}}\right) p(\boldsymbol{\theta}|y,\boldsymbol{x}) \mathrm{d}\boldsymbol{\theta} \tag{3}$$

There are some notes to be made here:

- Eq. (3) is derived under the assumption of independence of the observations.
- It would be possible to define a prior probability for $\underline{x}$. However, that would not affect the conditional problem (Gelman et al. 2013, Sect. 14.1).
- Explicit forms for the posterior parameter distribution and the predictive distribution exist in some cases. Still, in the vast majority of cases, those distributions should be simulated with Markov Chain Monte Carlo (MCMC) techniques.

A first consideration on the definition of the problem is related to time series forecasting applications. Indeed, those applications are based on the assumption of temporal dependence of variables; thus, the assumption of independence for deriving Eq. (3) would result in inferior predictive performances. For this specific case, the problem will be reformulated in Sect. 4.3. Some other questions arising directly from the above definitions are related to the assessment of the predictive performance of different models. The relevant theory will be introduced in Sect. 3.

2.2 Summary of review papers on probabilistic forecasting and prediction

Although the literature on predictions with machine learning is large, the same cannot be said for the topic of probabilistic predictions. A list of books and review papers on subtopics of probabilistic prediction can be found in Table 1. Taking a first look, none work has addressed all subtopics related to probabilistic prediction. Some papers are dedicated to probabilistic prediction with a narrow class of algorithms, such as distributional regression or deep learning. The topic of assessing the predictions is missing from those studies. Vice versa, papers related to scoring probabilistic predictions do not address the topic of algorithms. A large part of the literature examines the topic of forecasting, as well as that of model combinations.

Here, we intend to present a complete view of all topics along with their interplay. For instance, loss functions are an essential part of a machine learning algorithm; however, as can be seen from some recent review papers in deep learning (Abdar et al. 2021; Zhou et al. 2022) it is missing from the literature, which focuses on simulation-based techniques, either in simple simulation settings or in Bayesian ones.

Although some recent data competitions have highlighted the importance of machine learning algorithms in efficient forecasting, probabilistic forecasting review papers are relatively old and usually do not include the topic of machine learning. On the other hand, Bayesian textbooks focus on statistical considerations leaving less place to prediction considerations. Model combinations for probabilistic predictions with machine learning and probabilistic predictions focusing on extremes, as well as probabilistic predictions for spatial and spatio-temporal problems, are also missing from the literature.

3 Loss functions for assessing probabilistic predictions

The theory of loss (scoring) functions for point predictions has been extensively surveyed (see, e.g., Hyndman and Koehler 2006). Much of the development of machine learning algorithms is due to the knowledge of the properties of loss functions. Therefore, it is natural to ground our study on the theory of scoring functions and scoring rules for probabilistic predictions.

3.1 Assessment of quantile predictions

Machine learning practitioners are familiar with the use of loss functions to minimize the error between predicted and observed values for point prediction. Let $L(z, y)$ be a negatively oriented loss (scoring) function that returns a penalty when z is predicted and y realizes. When there are n observations $y_1, \ldots, y_n$ and n respective point predictions $z_1, \ldots, z_n$, issued by the regression algorithm then its performance can be scored by averaging the respective penalties:

$$S_n = (1/n) \sum_{i=1}^{n} L(z_i, y_i) \tag{4}$$

As a representative example, let $\rho_{1/2}(u)$ be half the absolute function

$$\rho_{1/2}(u) = |u|/2 \tag{5}$$

and

Table 1 Classification of review papers related to probabilistic forecasting and prediction

Reference	Title	Section(s)					
		3	4.1	4.3	4, 5	6	8
		Metrics—scoring	Bayesian modelling	Forecasting	Machine learning	Deep learning	Combinations
Genest and Zidek (1986)	Combining probability distributions: A critique and an annotated bibliography						✓
Chatfield (1993)	Calculating interval forecasts			✓			
Chatfield (1996)	Model uncertainty and forecast accuracy			✓			
Winkler (1996)	Scoring rules and the evaluation of probabilities	✓					
Hoeting et al. (1999)	Bayesian model averaging: A tutorial						✓
Tay and Wallis (2000)	Density forecasting: A survey			✓			
Lampinen and Vehtari (2001)	Bayesian approach for neural networks—review and case studies		✓			✓	
Geweke and Whiteman (2006)	Chapter 1 Bayesian Forecasting		✓	✓			
Gneiting and Raftery (2007)	Strictly proper scoring rules, prediction, and estimation	✓					
Robert (2007)	The Bayesian Choice		✓				
Casati et al. (2008)	Forecast verification: Current status and future directions	✓					
Khosravi et al. (2011)	Comprehensive review of neural network-based prediction intervals and new advances					✓	
Wilks (2011)	Chapter 8—Forecast Verification	✓					
Clarke and Clarke (2012)	Prediction in several conventional contexts		✓	✓	✓		
Vehtari and Ojanen (2012)	A survey of Bayesian predictive methods for model assessment, selection and comparison		✓				
Fahrmeir et al. (2013)	Regression: Models, Methods and Applications				✓		
Gelman et al. (2013)	Bayesian Data Analysis		✓				
Kneib (2013)	Beyond mean regression				✓		

Table 1 (continued)

Reference	Title	Section(s)					
		3	4.1	4.3	4, 5	6	8
		Metrics—scoring	Bayesian modelling	Forecasting	Machine learning	Deep learning	Combinations
Dawid and Musio (2014)	Theory and applications of proper scoring rules	✓					
Gneiting and Katzfuss (2014)	Probabilistic forecasting			✓			
Carvalho (2016)	An overview of applications of proper scoring rules	✓					
Koenker (2017)	Quantile regression: 40 years on				✓		
Rue et al. (2017)	Bayesian computing with INLA: A review		✓				
Fragoso et al. (2018)	Bayesian model averaging: A systematic review and conceptual classification						✓
Kabir et al. (2018)	Neural network-based uncertainty quantification: A survey of methodologies and applications					✓	
Ovadia et al. (2019)	Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift					✓	
Bassetti et al. (2020)	Density forecasting			✓			
Čížek and Sadıkoğlu (2020)	Robust nonparametric regression: A review				✓		
Wood (2020)	Inference and computation with generalized additive models and their extensions				✓		
Abdar et al. (2021)	A review of uncertainty quantification in deep learning: Techniques, applications and challenges		✓				
Bjerregård et al. (2021)	An introduction to multivariate probabilistic forecast evaluation			✓			
Hüllermeier and Waegeman (2021)	Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods				✓		
Kaplan (2021)	On the quantification of model uncertainty: A Bayesian perspective		✓				

Table 1 (continued)

Reference	Title	Section(s)					
		3	4.1	4.3	4, 5	6	8
		Metrics—scoring	Bayesian modelling	Forecasting	Machine learning	Deep learning	Combinations
Kneib et al. (2021)	Rage against the mean—A review of distributional regression approaches				✓		
Krüger et al. (2021)	Predictive inference based on Markov Chain Monte Carlo output		✓				
Makridakis et al. (2021)	The M5 uncertainty competition: Results, findings and conclusions			✓			
He et al. (2022)	Risk measures: Robustness, elicitability, and backtesting	✓					
Zhou et al. (2022)	A survey on epistemic (model) uncertainty in supervised learning: Recent advances and applications					✓	
Gawlikowski et al. (2023)	A survey of uncertainty in deep neural networks					✓	

$$L_{1/2}(z, y) = \rho_{1/2}(y-z) \tag{6}$$

be half the absolute error (AE) function. A machine learning algorithm trained by minimizing $L_{1/2}(z, y)$, predicts the median of the probability distribution of the response variable (Gneiting and Raftery 2007).

Similarly to predicting the median, one may receive a directive to predict a quantile of the probability distribution of the response variable. Let, $Q_{\underline{y}}(\alpha)$ defined by

$$Q_{\underline{y}}(\alpha) := F_{\underline{y}}^{-1}(\alpha), 0 \leq \alpha \leq 1 \tag{7}$$

be the αth quantile of $\underline{y}$. The quantile is a functional of the probability distribution. Intuitively, 100 α% of the $\underline{y}$ values are lower than $Q_{\underline{y}}(\alpha)$, while one could predict quantiles at a dense grid of quantile levels, to estimate the predictive distribution. A special case of Eq. (7) is the median of $\underline{y}$, $Q_{\underline{y}}(1/2)$.

Expanding the definition of the quantile to the regression setting, let $Q_{\underline{y}|\boldsymbol{x}}(\alpha|\boldsymbol{x})$

$$Q_{\underline{y}|\boldsymbol{x}}(\alpha|\boldsymbol{x}) := F_{\underline{y}|\boldsymbol{x}}^{-1}(\alpha) \tag{8}$$

be the αth quantile of $\underline{y}$ conditional on $\boldsymbol{x}$. Now, assume that one predicts the value z for $Q_{\underline{y}}(\alpha)$. Let also y be the respective realization of $\underline{y}$. In the familiar case of predicting the median of $\underline{y}$, the prediction's score would be $L_{1/2}(z, y)$ with optimal score being 0, when $z=y$. The absolute error function is generalized by the tilted absolute value function (Koenker and Bassett Jr 1978; Gneiting 2011a), defined by

$$\rho_{\alpha}(u) := u(\alpha-\mathbb{I}(u \leq 0)) \tag{9}$$

Here $\mathbb{I}(\bullet)$ denotes the indicator function and α is the quantile level of interest. For $\alpha=1/2$, Eq. (9) reduces to Eq. (5). The tilted absolute value function is positive and negatively oriented, i.e., the objective is to minimize it, and equals to 0, when $u=0$. Let

$$L_{\alpha}(z, y) := \rho_{\alpha}(y-z) \tag{10}$$

be the quantile loss function (or asymmetric piecewise linear scoring function) that returns the penalty $L_{\alpha}(z, y)$ to the prediction z of the αth quantile of $\underline{y}$ when y realizes (Gneiting 2011a).

While, intuitively, a point prediction z for the median of $\underline{y}$ would be considered satisfying if the penalty $L_{1/2}(z, y)$ is low, the extension to the quantile loss function is not straightforward. For understanding the properties of the quantile loss function, we illustrate in Fig. 1, $L_{\alpha}(z, 0)$ for varying predictive quantiles z at the quantile levels $\alpha \in \{0.05, 0.95\}$. The expectile loss function illustrated in the same figure will be explained later in Sect. 3.2. Assuming that $z>0$, it is clear that $L_{0.95}(z, 0) < L_{0.95}(-z, 0)$ and $L_{0.05}(z, 0) > L_{0.05}(-z, 0)$, while the quantile loss function is asymmetric. The asymmetry is due to assigning different weights to the error $y - z$ depending on its sign and allows estimating quantiles at levels different from ½, assuring that the correct ratio of observations 100α% lies below the prediction (Koenker and Hallock 2001). At level ½, $L_{1/2}(z, y)$ becomes symmetric; see Eq. (5).

A favourable property of the quantile loss function is that it is strictly consistent for the quantile of probability distributions with finite first moments (Murphy and Daan 1985; Gneiting 2011a). That means that following a directive to predict a functional of

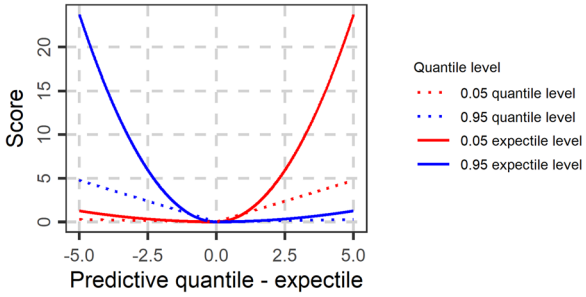

Fig. 1 Illustration of the quantile and expectile scores at the quantile levels $\alpha \in \{0.05, 0.95\}$ and expectile levels $\tau \in \{0.05, 0.95\}$, respectively, when $y = 0$ realizes and for varying predictive quantiles and expectiles z (see Eqs. (10) and (14))

the probability distribution (e.g., a quantile), a consistent scoring function will optimise the modeller's expected score (Ehm et al. 2016; Taggart 2022a).

In particular, a scoring function L: $D \times D \rightarrow [0, \infty)$, where D is the potential range of outcomes of a variable $\underline{y}$, is consistent for a functional T that maps $\mathcal{F} \rightarrow T(\mathcal{F}) \subseteq D$, relative to the class $\mathcal{F}$ of probability distributions if

$$\mathrm{E}_{\mathcal{F}} L(t, \underline{y}) \leq \mathrm{E}_{\mathcal{F}} L(z, \underline{y}) \tag{11}$$

for all probability distributions $F \in \mathcal{F}$, all $t \in T(\mathcal{F})$ and all $z \in D$. L is strictly consistent if it is consistent and equality of the expectations implies that $z \in T(\mathcal{F})$ (Gneiting 2011a). A functional is elicitable if there exists a scoring function that is strictly consistent for it (Gneiting 2011a).

It is proved that L is consistent for the αth quantile, if and only if is of the form

$$L_{\alpha}(z, y) = (g(y) - g(z))(\alpha - \mathbb{I}(y - z \leq 0)) \tag{12}$$

where g is a nondecreasing function on D. If g is strictly increasing, then L is strictly consistent. L is called generalized piecewise linear (GPL) of order α (Raiffa and Schlaifer 1961, p.196, Gneiting 2011a), while the quantile loss function $L_{\alpha}(z, y)$ is a special case of the GPL. Under the GPL, any αth quantile is an optimal point prediction (Gneiting 2011b). An intuitive explanation of the GPL scoring function is that g expresses some utility and that the loss by predicting z when y realizes is asymmetrically (depending on the sign of the difference) proportional to $|g(y) - g(z)|$ (Gneiting 2011b). Some further results on the interpretation of scoring functions that are consistent for quantiles as well as the use of plots for the comparisons of competing predictions can be found in Ehm et al. (2016).

3.2 Assessment of expectile predictions

Expectiles, introduced by Newey and Powell (1987), are functionals of the probability distribution. They are least squares analogues of quantiles, i.e., they are generalizations of the mean. Similar to Eq. (7), the expectile e_{τ} at the expectile level τ can be defined by inverting the following equation:

$$\tau = \mathrm{E}\left[\left|\underline{y} - e_{\tau}\right| \mathbb{I}\left(\underline{y} \leq e_{\tau}\right)\right] \Big/ \mathrm{E}\left[\left|\underline{y} - e_{\tau}\right|\right] \tag{13}$$

From Eq. (13) it is clear that e_{τ} is such that the mean distance from all $\underline{y}$ below e_{τ} is 100 τ% of the mean distance between $\underline{y}$ and e_{τ} (Daouia et al. 2018) and that $e_{1/2}$ is the mean of the probability distribution of $\underline{y}$.

The following asymmetric piecewise scoring function is strictly consistent for the expectile e_τ and has similar behaviour with the quantile loss function; see Fig. 1 (Gneiting 2011a).

$$L_\tau(z,y) := (y-z)^2|\tau-\mathbb{I}(y-z \le 0)| \tag{14}$$

When $\tau=½$, Eq. (14), reduces to half the squared error function

$$L_{1/2}(z,y) := \left((y-z)^2\right)/2 \tag{15}$$

It is proved that a scoring function is consistent for the τth expectile, if and only if is of the form

$$L_\tau(z,y) = \left(h(y)-h(z)-h'(z)(y-z)\right)|\tau-\mathbb{I}(y-z \le 0)| \tag{16}$$

where h is a convex function with subgradient $h\prime(z)$ (Gneiting 2011a). If h is strictly convex, then it is strictly consistent for the τth expectile. For some further results on the interpretation of scoring functions that are consistent for expectiles as well as the use of plots for the comparisons of competing predictions, as well as an interpretation of expectiles as a risk measure, see Ehm et al. (2016).

Expectiles are useful risk measures in econometrics, since they are both elicitable and coherent (quantiles are not coherent risk measures) (Bellini et al. 2014; Emmer et al. 2015). Furthermore, they can be informative regarding the size of losses, while quantiles are informative regarding the frequency of losses (Taylor 2021). Despite their favourable properties, expectiles are used less frequently compared to their counterparts (quantiles) possibly due to issues related with their interpretability (Waltrup et al. 2015). A relationship exists between quantiles and expectiles (Jones 1994). For very heavy-tailed distributions (with tail index larger than ½), extreme expectiles are larger than extreme quantiles in magnitude, while for probability distributions with tail index lower than ½, the opposite is true (Bellini et al. 2014).

For a generalization of quantiles and expectiles, see *M*-quantiles in Breckling and Chambers (1988), Koltchinskii (1997), with related consistent scoring functions (Huber quantile scoring functions) defined by Taggart (2022b) while also some intermediate loss functions between the quantile and the expectile losses exist, called *k*th power expectile losses (Jiang et al. 2021).

3.3 Proper scoring rules

Assume that a modeller predicts the full distribution $P \in \mathcal{P}$ for the response variable y, while F is the true predictive distribution. Let y be a future realization of y. Similarly to the point prediction case, the predictive distribution could be assessed based on the pair (P, y) (Dawid 1984). To assess the prediction, we define the scoring rule $S(P, y)$ (with negative orientation) that returns a score dependent on P and y, while the expected score under F is (Gneiting and Raftery 2007):

$$s(P,F) := \int S(P,y)\mathrm{dF(y)} \tag{17}$$

Note that the scoring rule resembles a scoring function, as an error measure, with difference being that its first argument is a distribution instead of a point. The scoring rule S is proper relative to the class of distributions $\mathcal{P}$ if

$$s(F,F) \leq s(P,F) \forall P,F \in \mathcal{P} \tag{18}$$

and strictly proper when $s(F, F) = s(P, F)$ if and only if $P = F$ (Winkler and Murphy 1968; Gneiting and Raftery 2007), i.e., a scoring rule is proper if it optimizes the expected score, when y from F realizes and the modeller issues a probabilistic prediction F (Gneiting and Raftery 2007). The divergence $d(P, F)$ is defined by:

$$d(P,F) := s(P,F) - s(F,F) \forall P,F \in \mathcal{P} \tag{19}$$

Interpreting $d(P, F)$ as a divergence is meaningful for strictly proper scoring rules (Bröcker 2009). In addition, the entropy $e(P)$ is defined by

$$e(P) := s(P,P) \forall P \in \mathcal{P} \tag{20}$$

The entropy $e(P)$ can be interpreted as lack of information (Bröcker 2009).

Minimizing the expected score, will lead to the true predictive distribution, regardless of the type of the scoring rule, although finite samples may affect the efficiency of the scoring rules (Loaiza-Maya et al. 2021).

3.3.1 Decomposition of proper scoring rules

A proper scoring rule can be decomposed into different components that contribute to the overall score of a probabilistic forecast, and these components can help us understand the properties of the probabilistic forecast. A general decomposition has been proposed by Bröcker (2009). Assuming that $\overline{\pi}$ is the probability distribution function of $\underline{y}$ (also termed climatology) and $\pi_{(\gamma)}$ is the conditional probability of $\underline{y}$ given the forecasting scheme γ, $E_{\mathcal{F}} S(\gamma, \underline{y})$ can be decomposed into three components:

$$E_{\mathcal{F}} S\left(\gamma, \underline{y}\right) = e\left(\overline{\pi}\right) - E_{\mathcal{F}} d\left(\overline{\pi}, \pi(\gamma)\right) + E_{\mathcal{F}} d(\gamma, \pi(\gamma)) \tag{21}$$

Here $e(\overline{\pi})$ is called the uncertainty of $\underline{y}$, $E_{\mathcal{F}} d(\overline{\pi}, \pi_{(\gamma)})$ is called resolution (or sharpness) and $E_{\mathcal{F}} d(\gamma, \pi_{(\gamma)})$ is called reliability.

The uncertainty is the expectation of the score of a climatology forecasting scheme. In the literature of point forecasting, the climatology is a forecasting scheme that is equal to the mean of dependent variable and can be used as the simplest benchmark. The resolution is a form of variance of $\pi_{(\gamma)}$. Larger resolution means a better score. We have a reliable forecast when $\gamma = \pi_{(\gamma)}$, therefore the reliability is the deviation of γ from $\pi_{(\gamma)}$.

3.3.2 Continuous ranked probability score (CRPS)

Probabilistic predictions should be sharp and well calibrated. Sharpness refers to the concentration of the distribution, with higher concentration being preferable. Calibration refers to the statistical consistency between the probabilistic predictions and the observations, in the sense that events that are predicted to occur with probability p, should be realized with frequency p. Sharpness is a property of the predictions while calibration is a property of both predictions and observations. Sharpness of the predictive distribution should be maximized subject to calibration, while a variety of proper scoring rules has been determined towards this direction (Gneiting et al. 2007).

Amongst the variety of proper scoring rules, the most widely used is the Continuous Ranked Probability Score (CRPS), defined by (Epstein 1969; Matheson and Winkler 1976; Gneiting and Raftery 2007)

$$\mathrm{CRPS}\,(F, y) := \int_{-\infty}^{\infty} (F(z) - \mathbb{I}(y - z \leq 0))^2 dz \tag{22}$$

Advantages of the CRPS is that is defined directly in terms of cumulative distribution functions thus it allows assessing forecasts in terms of samples (e.g., simulations) (Gneiting and Raftery 2007). Assuming that a simulation of K ensembles from a Bayesian statistical model, represents the predictive distribution, then the CRPS of the prediction can be written as a sum of quantile scores (see Eq. (10)) applied to each of the K ensemble members, where the level of kth lower ensemble member should be set equal to (k—0.5)/K (Bröcker 2012). Explicit expressions of the CRPS for given P and y or for ensemble (simulated) predictions as well as a related software implementation can be found in Jordan et al. (2019) and Krüger et al. (2021). For ordered ensemble members $z_{(i)}$ with $z_{(1)} < \ldots < z_{(K)}$ the CRPS can be estimated by

$$\mathrm{CRPS}\left(z_1, \ldots, z_k; y\right) = \left(2/K^2\right) \sum_{i=1}^{K} \left(z_{(i)} - y\right)\left(K\mathbb{I}\left(y - z_i < 0\right) - i + 1/2\right) \tag{23}$$

while the relevant expression for unordered ensembles in not recommended due to computational reasons.

Some further considerations from the theory of proper scoring rules follow. The CRPS can be decomposed into reliability, resolution and uncertainty (Hersbach 2000); see also Sect. 3.3.1, while probabilistic predictions for circular quantities can be can be assessed using a CRPS variant (Grimit et al. 2006). The CRPS generalizes the absolute error when the predicted distribution is a point measure, while an approximate relationship exists between the CRPS and the Root Mean Squared Error (RMSE, Leung et al. 2021).

3.3.3 Scoring rules for quantiles

As noted in Sect. 3.1, a probabilistic prediction can take the form of multiple predictive quantiles z_1, …, z_k at levels α_1, …, α_k. Then a proper scoring rule for assessing the predicted quantiles at the given levels is of the form

$$S\left(z_1, \ldots, z_k; y\right) = -\sum_{i=1}^{k} \left[\alpha_i s_i\left(z_i\right) + \left(s_i(y) - s_i\left(z_i\right)\right)\mathbb{I}\left(y - z_i \leq 0\right)\right] + h(y) \tag{24}$$

where $s_i(\bullet)$, $i = 1$, …, k is non-decreasing and h is an arbitrary function (Gneiting and Raftery 2007). Setting $k = 1$, $s_1(y) = y$ and $h(y) = \alpha\, y$ the quantile proper scoring rule reduces to the quantile scoring function. The quantile score can be decomposed in reliability, resolution and uncertainty (Bentzien and Friederichs 2014).

3.3.4 Scoring rules for intervals

Given quantile levels α_1 and α_2, with $\alpha_1 < \alpha_2$, a α_2—α_1 a prediction interval [q_1, q_2], corresponds to quantiles q_1 and q_2 of the response variable at levels α_1 and α_2, respectively.

A central prediction interval 1—α, corresponds to $\alpha_1 = \alpha/2$ and $\alpha_2 = 1—\alpha/2$. By setting $\alpha_1 = \alpha/2$, $\alpha_2 = 1—\alpha/2$, $s_1(y) = 2\ y/\alpha$, $s_2(y) = 2\ y/\alpha$, and $h(y) = 2\ y/a$ in Eq. (24), the central interval scoring rule is defined by (Dunsmore 1968; Winkler 1972; Gneiting and Raftery 2007)

$$S(z_1, z_2; y) = (z_2 - z_1) + (2/\alpha)(z_1 - y)\mathbb{I}(y - z_1 < 0) + (2/\alpha)(y - z_2)\mathbb{I}(z_2 - y < 0) \quad (25)$$

We note that we focus on prediction intervals for a single observation (Chatfield 1993). A possible requirement for a prediction interval is that it should have a specified coverage probability, i.e., a specified frequency of future observations falling within the prediction interval (Christoffersen 1998). Research has been done towards this direction (Chudý et al. 2020). Beyond considerations, on how this coverage probability should be set (Landon and Singpurwalla 2008), Askanazi et al. (2018) state that the problem of comparing prediction intervals when the coverage probability is specified but the quantiles are not, is difficult and perhaps unsolvable.

More generally, it has been shown that the shortest prediction interval and those intervals determined by an endpoint or midpoint are not elicitable, unless an endpoint is given via a quantile (Brehmer and Gneiting 2021; Fissler et al. 2021). Furthermore, the equal-tailed interval and the modal interval are elicitable (Brehmer and Gneiting 2021), while the mode and modal intervals are not indirectly elicitable (Dearborn and Frongillo 2020). Finally, it is possible to construct and evaluate expectile-bounded prediction intervals, similarly to quantile-bounded ones (Taylor 2021).

3.4 Some more proper scoring rules for assessing probabilistic predictions

A comprehensive list of proper scoring rules can be found in Gneiting and Raftery (2007) and Dawid and Musio (2014), including the log score (Good 1952), the Tsallis score (Tsallis 1988), the Brier score (Brier 1950), the Bregman score, the survival score, the Hyvärinen score (Hyvärinen and Dayan 2005), the composite score, the pseudo score, the diagonal score (Bouallègue et al. 2018), the energy score (Gneiting and Raftery 2007) and the variogram-based proper scoring rules (Scheuerer and Hamill 2015). Of special interest is the log score defined by

$$S(P, y) = -\log(p(y)) \quad (26)$$

The log score is the only proper scoring rule that is local, i.e., it depends on the predictive distribution only through its value at the event y that realizes (Gneiting and Raftery 2007; Dawid and Musio 2014), thereby ignoring probabilities of events that could have happened but did not (Krüger et al. 2021); see also proper local scoring rules of order k in Ehm and Gneiting (2012) and Parry et al. (2012). The log score is connected with likelihood inference in statistical modelling. In comparison, for instance, with the CRPS, a major drawback of the log score is that is restricted to predictive densities (Gneiting and Raftery 2007). Still, there are also advantages of local proper scoring rules. For such advantages, see the discussion in Du (2021).

Some other proper scoring rules are the error-spread scoring rule, which is formulated with respect to moments of the predictive distribution (Christensen et al. 2015; Christensen 2015), a scoring rule for simultaneous events (Grant et al. 2019), a proper scoring rule motivated by the form of Anderson–Darling distance of distribution functions (Barczy 2022), threshold-and quantile-weighted scoring rules (Gneiting and

Ranjan 2011), scoring rules that exploit temporal dependence between events (Lai et al. 2011), joint scoring rules (Lichtendahl and Winkler 2007) and the asymmetric continuous probabilistic score (ACPS) (Iacopini et al. 2022).

3.5 Some more related concepts

Most score values are scale depended (i.e., they depend on the magnitude of observations), therefore skill scores may be used instead to compare predictions for multiple cases (e.g., for multiple time series). Skill scores are standardized version of scoring rules, but they are not necessarily proper (Gneiting and Raftery 2007).

Arbitrary scoring rules can be made proper, while details on such constructions (called properizations) can be found in Dawid (2007) and Brehmer and Gneiting (2020). Tailored scoring rules to client's specified requirements should be selected among the list of scoring rules presented in the previous sections, while it is possible to construct new scoring rules according to requirements of prediction users; see, e.g., Johnstone et al. (2011), Machete (2013) and Merkle and Steyvers (2013).

Finally, Liu et al. (2020) argue that minimizing a proper scoring rule over the input space is not possible, because the true distribution of data cannot be learned by a model, therefore they proposed minimizing a scoring rule with respect to all possible true distributions, transforming the minimization to a minimax problem.

In economic and financial sciences, backtesting is used to assess the performance of a model when predicting a time series. For example, when predicting Value at Risk (VaR), which is a quantile of the predictive distribution, one is interested in ensuring that the conditional coverage of predictions is equal to the nominal level of VaR. Misspecification of the conditional coverage is termed as a violation. A review of backtesting methods for VaR can be found in Zhang and Nadarajah (2018) who have grouped the backtesting methods into four categories: unconditional test methods (e.g., Probability of Exceedance (POF) test by Kupiec 1995), conditional test methods, independence property test methods (e.g., the dynamic quantile (DQ) test by Engle and Manganelli 2004) and other test approaches. Unconditional methods focus on the unconditional coverages, while independence property test methods evaluate the extent to which a VaR measure's performance is independent from one period to the next (Zhang and Nadarajah 2018). The concept of elicitability presented earlier regards comparison of multiple models, while backtesting refers to the validation of a single model (He et al. 2022).

4 Early history and simple models

Before proceeding to machine learning models, a small overview of simpler models that are appropriate for probabilistic predictions may serve in understanding how more complex and accurate models can be built. In particular, we will present the theory of the simple Bayesian statistical ordinary linear regression model, the linear in parameters quantile regression, as well as an overview of time series and copula models. Much of the theory of machine learning models is based on simpler models, with added complexity resulting in improved performance.

4.1 Bayesian statistical models

Returning to the Bayesian statistical models and following Eq. (3), it is proved that the predictive distribution for the Gaussian ordinary linear regression model with unknown parameters, when assigned a uniform noninformative prior distribution to (β, logσ), where σ is the variance of the error term ε_i

$$y_i = \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \varepsilon_i, i = 1, \ldots, n \tag{27}$$

is multivariate Student (Gelman et al. 2013, Chapter 14) with n—p degrees of freedom, location parameter $\widetilde{\boldsymbol{X}}\, u$ and squared scale matrix $s^2\,(\boldsymbol{I} + \widetilde{\boldsymbol{X}}\, \boldsymbol{V}_u\, \widetilde{\boldsymbol{X}}^{\mathrm{T}})$, where $\widetilde{\boldsymbol{X}}$ is the matrix of future $\widetilde{\boldsymbol{x}}$ values, $\boldsymbol{X}$ is the $n \times p$ matrix of predictors (recall from Sect. 2.1 that n is the number of samples), $\boldsymbol{y}$ is the $n \times 1$ vector of responses and

$$u = \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T y \tag{28}$$

$$\boldsymbol{V}_u = \left(\boldsymbol{X}^{\mathrm{T}} \boldsymbol{X}\right) \tag{29}$$

$$s^2 = (1/(n-k))\,(\boldsymbol{y}-\boldsymbol{X}u)^{\mathrm{T}}(\boldsymbol{y}-\boldsymbol{X}u) \tag{30}$$

Obviously, it is possible to estimate the predictive distribution for the linear regression, with MCMC simulations (Gelman et al. 2013, Chapter 14). Early considerations of predictive inference for Gaussian models include Geisser (1965), Chew (1966) and Eaton et al. (1996), among others. Predictive distributions for Bayesian statistical models beyond the Gaussian case may not be defined explicitly; therefore, simulations in a MCMC setting may be appropriate and perhaps the only means to proceed.

Although Bayesian statistical modelling is mostly concerned with in-sample performances, where it is essential to specify correctly the data generating process, cross-validation (with out-of-sample performance) using various loss functions has also been the subject of recent research. For relevant examples, see Vehtari and Ojanen (2012), Piironen and Vehtari (2017) and Vehtari et al. (2017). Problems related to cross-validation schemes are that Bayesian statistical models are mostly applied to small datasets, while as the dataset increases, the computational cost of cross-validation may also increase significantly.

To overcome computational costs related to the increasing size of datasets and intractability of likelihoods, Approximate Bayesian Computation (ABC) has been proposed as a solution. ABC is an algorithm that simulates artificial data using parameters from the prior distribution and keeps the data that are close to the observed data according to a distance metric. ABC methods for estimating predictive distributions has been surveyed by Frazier et al. (2019) and has been found to provide efficient predictions that are nearly identical, although inferior, to exact ones.

Focused Bayesian prediction (Loaiza-Maya et al. 2021) is a method for Bayesian predictions, in which correct specification of the data generation process is not required. Updating is done using a specified measure (e.g., a proper scoring rule), while the method is also applicable to improve combinations of algorithms (see also Sect. 8 for an exposition of combinations of algorithms).

4.2 Quantile and expectile regression linear in parameters

4.2.1 Quantile regression

As mentioned in Sect. 3.1, quantile predictions can be assessed using the quantile loss function defined by Eq. (10). Therefore, the idea to estimate the parameters of the linear regression model defined by Eq. (27), elaborated by Koenker and Bassett Jr (1978), is natural when one's model is intended to predict quantiles. Optimization of the model is done by averaging the losses over the sample, similarly to what is made for least squares regression. Quantile regression is appropriate when (Waldmann 2018): (a) one is interested in predicting functionals beyond the mean; (b) the parametric form of the predictive distribution is now known; (c) outliers exist (in which case quantile regression is robust); and (d) in presence of heteroscedasticity (i.e., when the variance depends on the covariates). Several developments of quantile regression have been proposed. Although we cannot be exhaustive on them, the following may be of interest, while a brief exposition can be found in Koenker (2017) and an extensive treatment can be found in Koenker et al. (2017).

Regularization in quantile regression: In cases where the estimation of the model's parameters is unstable, a regularization process can be applied to improve inference (Bickel and Li 2006). This can be made, for instance, by using ridge regression (Hoerl and Kennard 1970) or lasso (Tibshirani 1996), among other methods for achieving prediction improvements, which may include models with many parameters or high dimensional problems. Regularization has been implemented in quantile regression models, as it is described in overviews such as those by Mizera (2017), Wang (2017) and Belloni et al. (2017). A characteristic example of such quantile regression models is the lasso estimator by Ye and Padilla (2021).

Bayesian statistical modelling and quantile regression: Bayesian statistical models for quantile regression can be useful in cases where inference on the parameters is required in the usual Bayesian settings, or in cases where the optimization procedure is difficult. The specification of parametric likelihoods is not possible in quantile regression methods; thus, working likelihoods are usually implemented. An overview of methods on Bayesian quantile regression can be found in Wang and Yang (2017), while the first relevant study appears to be Yu and Moyeed (2001).

Multivariate quantiles: Quantile regression can be extended to the multiple-output case. Assuming that quantiles of $d \geq 2$ variables have to be predicted simultaneously, the extensions are natural, although the problem's complexity increases considerably. An overview of methods can be found in Hallin and Šiman (2017). Approaches to the problem include the directional one, in which the multivariate problem is reduced to several univariate ones, in which the focus is on the marginal distributions (see, e.g., Hallin et al. 2010). Another approach is the direct one, with the following alternatives: (α) Extension of the loss function to cover the spatial case (spatial loss function); (b) the concept of elliptical quantiles; and (c) the concept of depth-based quantiles. An example for estimating simultaneously multiple quantiles is proposed by Firpo et al. (2021) using a generalized method of moments (GMM).

Quantile crossing problem: In the usual case, conditional quantiles in quantile regression are estimated independently, i.e., by fitting different models at varying quantile levels. Consequently, it is possible that a predicted quantile at a given level may be higher than a predicted quantile at a higher level. That problem of quantile regression algorithms

is called quantile crossing. Several methods have been proposed to remedy the quantile crossing problem,, e.g., by sorting the estimated quantiles (Chernozhukov et al. 2010) or by enforcing non-crossing constraints (Takeuchi et al. 2006).

Survival analysis with quantile regression: In survival analysis, the duration until one event occurs is of interest. Censoring in survival analysis refers to the case that the time to event cannot be observed, for example, because the study may stop before the event realizes. Overviews of methods for survival analysis including the censoring case can be found in Ying and Sit (2017), Li and Peng (2017) and Peng (2017, 2021), while the first relevant study appears to be Powell (1986).

Semiparametric models: Quantile regression for semiparametric models (i.e., regression models that combine parametric and nonparametric models) has been proposed by Waldmann et al. (2013). Components of the models may include nonlinear effects, spatial effects, non-normal random effects and more. Due to optimization difficulties, such models are formulated in Bayesian settings.

Panel analysis: Quantile regression has been applied to panel data. For instance, Geraci and Bottai (2014) modelled multiple random effects.

Optimization procedures: An important problem of quantile regression algorithms is that the quantile loss function is not everywhere differentiable; therefore, gradient based optimization methods are not always applicable. For efficient achieving optimizations, approximations to the quantile loss function have been employed. A relevant example can be found in Zheng (2011).

4.2.2 Expectile regression

Similarly to what applies to quantile regression, the parameters of the linear regression model, defined by Eq. (27), can be estimated by minimizing the expectile loss function defined by Eq. (14). Expectile regression has been elaborated by Newey and Powell (1987) and has similar properties with quantile regression when contrasted to Bayesian statistical models. On the other hand, there are some differences between quantile and expectile regression, due to properties of the respective loss functions (see Sect. 3.2). The progress in expectile regression models follows that on quantile regression, albeit at a slower pace. Perhaps this can be attributed to issues related to the interpretability of expectiles, as already explained in Sect. 3.2 (Waltrup et al. 2015).

Naturally, similar themes examined in quantile regression, find their counterparts in expectile regression. To mention some, one can find semiparametric models (Sobotka and Kneib 2012; Sobotka et al. 2013; Spiegel et al. 2017), regularization (Waldmann et al. 2017; Zhao and Zhang 2018; Zhao et al. 2018; Liao et al. 2019), Bayesian statistical modelling (Waldmann et al. 2017) and survival analysis (Seipp et al. 2021).

4.3 Forecasting with time series models

Let $\{y_1, \ldots, y_t\}$ be a time series indexed by $i=1, \ldots, t$, $t \geq 1$. A time series forecasting problem is defined as the prediction of a future variable $\underline{y}_{t+h}$, at forecasting horizon $h>0$, conditional on observations $\{y_i\}$, $i \leq t$. One-step ahead forecasting is defined for $h=1$ and multi-step ahead forecasting is defined for $h>1$. When the probability distribution of $\underline{y}_{t+h}$ is of interest, we have a probabilistic forecasting problem. The term automatic forecasting, that describes systems that forecast time series given only observed input and required in

the case of large number of time series, extends also in the case of probabilistic forecasting (Ord and Lowe 1996).

Compared to IID data modelled by traditional regression models, time series data contain additional information due to temporal dependence; therefore, models may benefit by incorporating such information. A category of such models are stochastic processes, defined by $\{\underline{y}_1, \ldots, \underline{y}_t\}$. In the usual Bayesian setting, incorporating temporal dependence complicates things a little, in the sense that Eq. (3) should be corrected to exploit such information. Practically, the one-step ahead forecasting problem is defined after appropriate modifications of Eq. (3), by

$$p(\underline{y}_{t+1}|y_1, \ldots, y_t) = \int p(y_{t+1}|\boldsymbol{\theta}, y_1, \ldots, y_t) p(\boldsymbol{\theta}|y_1, \ldots, y_t) \mathrm{d}\boldsymbol{\theta} \tag{31}$$

The likelihood function, used to estimate the posterior distribution of $\boldsymbol{\theta}$, is also more complicated since the dependence between the variables of the stochastic process should be modelled.

4.3.1 Bayesian forecasting models

Bayesian statistical modelling constitutes a formal means for estimating the predictive distribution of future variables in time series forecasting problems. Due to the complicated nature of the parameters' likelihood function, it is not always possible to derive explicit forms of the predictive distribution using non-informative distributions, in contrary to regression problems (see, e.g., Sect. 2.1). The formal theory for probabilistic forecasting of time series using Bayesian modelling and stochastic processes can be found in the overview by Geweke and Whiteman (2006). Some concepts and methods that are representative of what can be met in practice follow:

Autoregressive moving average (ARMA) models: Bayesian modelling of ARMA models (that also include the special cases of AR and MA models) using conjugate priors and simultaneously estimating the required number of parameters has been done by Monahan (1983).

Autoregressive fractionally integrated moving average (ARFIMA) models: Bayesian modelling of ARFIMA models using non-informative priors has been done by Pai and Ravishanker (1996), while Durham et al. (2019) propose solutions for large time series.

Autoregressive with exogenous variables (ARX) models: Multi-step ahead probabilistic forecasting for ARX models haw been studied by Liu (1994), who implemented a normal-gamma prior parameter distribution, and by Liu (1995) who implemented conjugate prior distributions in an ARX model with random exogenous variables.

Vector models: Vector autoregressive (VAR) models have been studied by Sims and Zha (1998) who used informative priors, and vector ARFIMA (VARFIMA) models have been studied by Ravishanker and Ray (2002). Bayesian modelling of global vector autoregressive models (B-GVAR) has been examined by Cuaresma et al. (2016).

Exponential smoothing models: Exponential smoothing models (including their variants,, e.g., the Holt-Winters model) are statistical models specialized for forecasting. Some variants of exponential smoothing models may be subcases of ARMA models, but are referred as a different category, due to their success in practical applications. Bayesian settings for such models have been proposed,, e.g., by Bermúdez et al. (2010) for the

Holt-Winters model, and by Bermúdez et al. (2009) for the multivariate Holt-Winters model.

Non-linear time series models: AR models with time varying parameters (i.e., a class of non-linear time series model) has been examined in Bayesian settings by Müller et al. (1997).

Generalized autoregressive conditional heteroscedasticity (GARCH) time series models: Bayesian inference for GARCH time series models and its variant EGARCH (exponential GARCH) has been done by Vrontos et al. (2000). Bayesian inference methods for a variety of GARCH models, including, e.g., the Glosten–Jagannathan–Runkle (GJR-ARCH) model have been reviewed by Virbickaite et al. (2015).

Approximate Bayesian forecasting: In time series forecasting problems, the size of the length of the time series may prohibit Bayesian methods from being practically applicable. As already noted in Sect. 4.1, approximate Bayesian computation may be a benefitting solution with little cost in predictive performance (Frazier et al. 2019).

4.3.2 Bootstrap-based forecasting models

Bayesian statistical models are parametric and may not be preferable in many cases; see related discussion in Sect. 10. An alternative method to obtain probabilistic forecasts with time series models is resampling (or bootstrapping) techniques (Efron 1979). The distribution of the innovations of the time series model (innovation is the random component of the model that essentially models its error) can be approximated by resampling the residuals of the fitted model. A first overview of such models can be found in Chatfield (1993). Some more recent resent developments include:

Autoregressive moving average (ARMA) models: The case of highly persistent AR models has been studied by Clements and Kim (2007).

Vector models: Bootstrap-based probabilistic forecasts have been studied by Grigoletto (2005). The case of multi-step ahead probabilistic forecasts has been investigating by Staszewska-Bystrova (2011), Fresoli et al. (2015) and Fresoli (2022).

State-space models: The case of periodic state-space models has been investigated by Guerbyenne and Hamdi (2015).

Non-linear time series models: Self-exciting threshold autoregressive (SETAR) model have been examined by Li (2011).

4.3.3 Quantile regression-based models

Quantile regression is an alternative to Bayesian statistical models for probabilistic predictions of regression models; see Sect. 4.2.1. Therefore, it is natural to expand such modelling techniques to time series models. Time series models, introduced in Sect. 4.3.1 and 4.3.2, have their quantile loss based counterparts, as follows:

Autoregressive moving average (ARMA) models: AR models fitted with the quantile loss function have been investigated by Koenker and Xiao (2006). Quantile crossing issues have been addressed by Gourieroux and Jasiak (2008). A development related to the usual fitting procedure of ARMA models that considers correlations has been investigated in the context of quantile-based models by Li et al. (2015).

Conditional Autoregressive Value at Risk (CAViaR) models: CAViarR models predict quantiles by specifying their evolution over time using an autoregressive process (Engle and Manganelli 2004).

Conditional Autoregressive Expectile (CARE) models: CARE models are similar to CAViaR models, with their difference being that they are estimated using an expectile loss function (Kuan et al. 2009).

Exponential smoothing models: Quantile regression for exponential smoothing models has been studied by Taylor and Bunn (1999).

Generalized autoregressive conditional heteroscedasticity (GARCH) time series models: GARCH models have been investigated by Xiao and Koenker (2009).

4.3.4 Other forecasting models

Alternative practices for probabilistic forecasting with time series models are presented in Chatfield (1993) and include a Gaussian analytical approximation of the model's errors or approximation of the variance of the model's error. Progress in alternative techniques is also of interest, since relevant ideas may be applicable to general machine learning frameworks. Some examples include probabilistic forecasts of VAR models using computationally efficient algorithms for computing multi-dimensional integrals involving the Gaussian density function (Chan 1999), Gaussian approximations (De Luna 2000), first-order Taylor approximations (Snyder et al. 2001), approximations based on asymptotic properties (Hansen 2006) and approximation of the error's distribution (Wu 2012; Lee and Scholtes 2014). The problem of multi-step ahead probabilistic forecasting has been too complex to solve (Regnier 2018).

4.4 Copula-based regression models

Copulas are multivariate cumulative distribution functions with uniform marginal probability distributions for each variable (Nelsen 2006). The importance of copulas arises from Sklar's theorem (Sklar 1959), according to which, under certain conditions, there exists a unique copula function C, such that

$$H(x, y) = C(F(x), G(y)), \forall x, y \tag{32}$$

where H is the joint distribution function of x and y with respective marginal distributions F and G. The converse is true, i.e., if C is a copula and F and G are distribution functions, then H defined by Eq. (32) is a joint distribution function with marginals F and G.

Copula-based regression models have been introduced relatively very recently; however, due to their resemblance to simpler statistical models, we prefer to introduce them in this Section. In practice, copula-based regression models work by substituting $F(x)$ with the marginal distributions $\boldsymbol{F(x)} = (F_1(x_1), \ldots, F_p(x_p))$ of the predictor variables. The idea of copula-based regression is to obtain the distribution G, conditional on $\boldsymbol{x}$ and specified C (that models the dependence between variables) (Noh et al. 2013). An early exploitation of the idea can be found in Sungur (2005). Copulas are parametric models; therefore, the key is to specify correctly the copula function. Since copulas are parametric models, extensions to probabilistic predictions for the regression models are natural. Furthermore, the dependence between variables is modelled separately from the marginal distributions, thus facilitating the modelling procedure. In the following, we will present several version based on combinations of components presented earlier.

Estimation procedures: Noh et al. (2013) consider a semiparametric approach for estimating the model. In particular, they propose parametric modelling of the copula function and non-parametric modelling of the marginal distributions. However, when

the copula function is misspecified, the results may be largely unreliable, because the regression function is monotone in the predictor variables (Dette et al. 2014).

Direct quantile estimation: Quantile estimation by using directly the estimated conditional distribution functions has been examined by Kraus and Czado (2017) using vine copulas; see also Größer and Okhrin (2021) for the usefulness of using this class of copulas. Their method was extended to improve the conditional likelihood by Tepegjozova et al. (2022).

Quantile regression: Copula-based regression has been extended to the quantile regression case, both for IID data and for time series by Noh et al. (2015). The quantile regression case is also examined by Rémillard et al. (2017).

Survival analysis with copula-based regression: Copula-based regression for estimating quantiles in survival analysis has been investigated by De Backer et al. (2017).

Regularization in copula-based regression: High-dimensional Gaussian copula-based regression has been investigated by Tony Cai and Zhang (2018). He et al. (2018) propose transforming the variable selection problem to a multiple testing one, thus reducing the number of redundant variables due to the stopping rule.

Bayesian statistical modelling: Bayesian statistical modelling for copula-based regression models has been proposed by Klein and Kneib (2016), who also add non-linear predictors for the parameters of the model.

Various copulas: Among various types of copulas, some authors focus on a specific class, that may have some merits in particular situations. These include, for instance, Gaussian copula regression (Masarotto and Varin 2017; Zhao et al. 2020) and vine copulas for handling mixed (continuous and discrete) data, conditional heteroscedasticity, large-scale, nonlinear and non-Gaussian data (Chang and Joe 2019; Liu and Li 2022). Several other models are surveyed by Kolev and Paiva (2009).

Missing data: The case of missing observations has been investigated by Hamori et al. (2020) using a calibration estimation approach.

5 Machine learning algorithms

Now that fundamental concepts and simpler models have been introduced, it is time to continue with machine learning models for probabilistic predictions. A description of most machine learning models can be found in textbooks, such as those by Hastie et al. (2009), while we consider that the reader is familiar with machine learning algorithms (e.g., with random forests or boosting algorithms) in the following. Transforming a machine learning model to issue probabilistic predictions is possible by combining relevant fundamental concepts introduced earlier (see also Mukhopadhyay and Wang 2020, for a relative view that focuses on the connections between parametric and non-parametric models).

5.1 Gaussian process regression

A Gaussian process is a stochastic process, with any finite number of the process variables having a joint Gaussian distribution (Rasmussen 2004). Due to the multivariate Gaussian property, the predictive distribution can be estimated in Bayesian settings similarly to Gaussian models (e.g., linear models as shown in Sect. 4.1 or forecasting models as shown in Sect. 4.3.1). The form of the predictive distribution for various cases of prior distributions can be found, for example, in Rasmussen and

Williams (2006) and Jankowiak et al. (2020) or in Girard et al. (2003) when the scope is time series forecasting and Corani et al. (2021) when the scope is automatic time series forecasting.

Probabilistic predictions with nonstationary Gaussian process models have been proposed by Liang and Lee (2019), while the relevant case of high-dimensionality has been investigated by Risser and Turek (2020). Heteroscedasticity has been modelled with Gaussian processes by Binois et al. (2018) and Binois and Gramacy (2021).

5.2 Generalized additive models for location, scale and shape (GAMLSS)

Beyond Bayesian statistical and quantile regression models, another category is that of generalized additive models for location, scale and shape (GAMLSS, Rigby and Stasinopoulos 2005). GAMLSS are an extension of the more restrictive generalized additive models (GAMS, Hastie and Tibshirani 1986) and model directly the parameters of the predictive distribution as functions of the predictor variables in regression settings; therefore, the full predictive distribution is estimated, unlike quantile regression models that estimate functionals of the predictive distribution. For this reason, GAMLSS have also been termed as distributional regression models (Umlauf and Kneib 2018). The models are general in the sense that any probability distribution can be modelled, thus moving beyond the Gaussian one, and including parameters beyond the mean and variance (e.g., the skewness). The parameters can be modelled by parametric, as well as additive non-parametric functions (e.g., smoothing splines, penalized splines, local regression smoothers) of the predictor variables, while random effects (e.g., spatial random effects) can also be included as well as combinations of components. The model is trained by penalized likelihood optimization (see also Sect. 3.4, for the connection of likelihood with the log score). Therefore, GAMLSS models need specification of the form of the dependent variable's probability distribution function, the link function and the estimation procedure (Henzi et al. 2021b). Software implementation of GAMLSS models can be found in Stasinopoulos and Rigby (2007). The combination of earlier ideas that led to the development of GAMLSS can be found in Green (2013). Various improvements and variants of GAMLSS models have also appeared (many are also summarized by Stasinopoulos et al. 2018). Those include:

Alternative distributions: Modelling of distributions with four parameters has been proposed by Rigby and Stasinopoulos (2006). Extensive investigation of symmetric probability distribution functions has been done by Ibacache-Pulgar et al. (2013).

Copulas: GAMLSS is extended to the case of predicting multiple responses modelled by copulas (Marra and Radice 2017) and to the case of predicting multiple instances of a single variable by modelling their dependencies using copulas (Smith and Klein 2021).

Bayesian statistical modelling: An extension of GAMLSS, termed Bayesian Additive Models for Location, Scale, and Shape (BAMLSS), that includes Bayesian statistical modelling has been proposed by Umlauf et al. (2018, 2021) and Umlauf and Kneib (2018). These authors bring Bayesian-based benefits related to efficient estimation procedure and statistical inference. Bayesian multivariate modelling (i.e., prediction of distributions of multiple variables simultaneously) has been investigated by Klein et al. (2015a). The Poisson model for an excess number of zeros has been especially investigated by Klein et al. (2015b) in the previous context. Extension to the case of skew distributions has been proposed by Michaelis et al. (2018).

Other models: The case of using instrumental variables in distributional regression has been investigated by Briseño Sanchez et al. (2020). Efficient estimation of censored or truncated dependent variables has been implemented by Messner et al. (2016).

5.3 Random forests

Random forests are ensembles of decision trees that are built using bootstrap aggregation with some additional randomization (Breiman 2001a). Quantile regression forests (Meinshausen 2006) are a generalization of random forests that can estimate conditional quantiles in a non-parametric way. Trees in quantile random forests are grown similarly to random forests and the conditional distribution is again estimated using the tree responses, although instead of the average of the ensembles, an indicator function is used to define quantiles. Quantile regression forests are consistent for quantile estimation under certain general conditions. Several other variants and improvements of quantile regression forests have been proposed, as presented in the following (see also the comparison of variants by Roy and Larocque 2020):

Bias correction: Quantile regression forests prediction bias is corrected in a post-processing framework in which the error of the predictions is subtracted at a second stage (Tung et al. 2014; Nguyen et al. 2015).

Splitting rule: In decision trees, the splitting rule (e.g., rules based on Gini impurity, variance reduction) is of great importance. Bhat et al. (2015) proposed a splitting rule based on the quantile loss function that resembles quantile regression. Generalized random forests are based on a gradient-based splitting scheme to optimize heterogeneity (Athey et al. 2019). Local linear forests model smooth signals and fix boundary bias issues (Friedberg et al. 2020).

Distributional regression: Distributional regression for random forests has been proposed by Schlosser et al. (2019) in similar way to GAMLSS, but using decision trees as base learners. An application on circular data using the von Mises distribution has been demonstrated by Lang et al. (2020). Modelling of the Beta distribution, which is particularly useful for variables bounded in (0, 1) has been proposed by Weinhold et al. (2020) using a tailored likelihood based splitting rule.

Other models: An online learning variant of quantile regression forests has been developed by Vasiloudis et al. (2019). Out-of-bag (OOB) prediction intervals have been proposed by Zhang et al. (2020). With this approach, a single predictive distribution is estimated, instead of separate conditional distributions for new cases. Random forests tailored for time series data (which have some additional useful properties related to temporal dependence; see Sect. 4.3) have been developed by Davis and Nielsen (2020) and may have potential for probabilistic prediction. The framework of Lu and Hardin (2021) for estimation of prediction intervals is based on OOB weighting of OOB prediction errors.

5.4 Boosting algorithms

Boosting algorithms are ensemble learning algorithms, in which the base learners are trained sequentially to minimize a loss function (Friedman 2001). Base learners are added in the ensemble with the aim to correct the errors of previous base learners. Two key components of boosting algorithms are the loss function that can be tailored to user needs and the type of base learners, in the sense that multiple types of base learners can be combined in an ensemble. Furthermore, boosting algorithms include an intrinsic

mechanism for variable selection that is particularly suited for high-dimensional data (Mayr et al. 2014a, 2014b, 2017; Mayr and Hofner 2018). Friedman's (2001) original algorithm can optimize the quantile loss function, for example, by using decision trees as base learners. Therefore, it can predict conditional quantiles. Several other variants of boosting can give probabilistic predictions, based on its properties. In the following, we present some of them:

Quantile regression: Boosting for quantile regression has been proposed by Hofner et al. (2014) using algorithms particularly suited for high–dimensional data and including a diverse set of base learners ranging from linear effects to splines, spatial effects, random effects and more. Boosting algorithms that have shown exceptional performance in practical situations, with characteristic examples of those being the XGBoost (Chen and Guestrin 2016) and LightGBM (Ke et al. 2017), can also be optimized with the quantile loss function.

Expectile regression: Expectile regression with boosting has also implemented in the above framework by Hofner et al. (2014).

Distributional regression: Distributional regression with boosting using diverse base learners (ranging from linear effects to splines, spatial effects, random effects and more) has been proposed by Mayr et al. (2012), while the respective software implementation can be found in Hofner et al. (2016). The algorithm is termed gamboostLSS, for boosting generalized additive models for location, scale and shape. Modelling of the Beta distribution for bounded dependent variables has been investigated by Schmid et al. (2013). Stability selection for including the important predictors in the boosting ensemble has been proposed by Thomas et al. (2018), while a technique to deselect unimportant base learners has been proposed by Strömer et al. (2022). Natural Gradient Boosting (NGBoost) for probabilistic prediction (Duan et al. 2020) is another boosting-based distributional regression algorithm that has been implemented with multiple proper scoring rules.

Other models: Random gradient boosting (Yuan 2015) is an algorithm that combines random forests and boosting for exploiting advantages from both techniques for probabilistic prediction. Contrast trees can be used to find possible lack of fit for a model, while models can be improved by applying contrast boosting which can also provide probabilistic predictions (Friedman 2020). Probabilistic gradient boosting machine (PGBM) are used to estimate the mean and variance, which can then be used to sample from a specified distribution (Sprangers et al. 2021).

5.5 A list of other machine learning models

Beyond the main classes of models presented in Sects. 5.1–5.4, other classes of models less frequently used are presented in the following.

5.5.1 Bayesian Additive Regression Trees (BART)

Bayesian Additive Regression Trees (BART) are a sum of decision trees in which a regularization prior is imposed on the parameters of the model (Chipman et al. 2010). BART are motivated by boosting algorithms, while they allow for probabilistic predictions. A variant of BART based on multiplicative regression trees allows modelling heteroscedasticity (Pratola et al. 2020).

5.5.2 Transformation models

Transformation models are practically obtained by inverting the regression equation. An extension of transformation models for estimating predictive distributions has been proposed by Hothorn et al. (2014). In contrast to GAMLSS, conditional transformation models rely on less assumptions (Hothorn et al. 2014). Investigation of maximum likelihood estimators for conditional transformation models has been done by Hothorn et al. (2018), while the relevant software implementation can be found in Hothorn (2020a). Boosting-based training of conditional transformation models has been extensively investigated by Hothorn (2020b), while decision trees motivated implementation has been proposed by Hothorn and Zeileis (2021). Autoregressive transformation models (that combine properties of time series models and transformation models) have been introduced by Rügamer et al. (2023a).

5.5.3 Conformal prediction

Conformal prediction is a technique to estimate marginal prediction intervals that include the dependent variable with a specified frequency (Vovk et al. 2005; Shafer and Vovk 2008; Gammerman 2012). The algorithm is based on a nonconformity measure (e.g., one that measures some kind of distance) for a given significance level and includes predictions conditional on the values of the measure using p values. The algorithm can be combined with any machine learning algorithm (e.g., those presented earlier). The method seems to be different from the example of proper scoring rules, while as also stated in Sect. 3.3.4, it is difficult to compare intervals with specified coverage probability.

Conformal regression with various machine learning algorithms: Conformal prediction with k-nearest neighbours regression (k-NN) combined with the introduction of six nonconformity measures has been investigated by Papadopoulos et al. (2011). Conformal prediction with k-NN for time series forecasting has been investigated by Tajmouati et al. (2022). Conformal prediction with random forests has been presented by Johansson et al. (2014), with survival random forests by Bostrom et al. (2017) and with Mondrian trees by Johansson et al. (2018).

Conformal prediction with quantile regression algorithms: Conformalized quantile regression combines conformal prediction with quantile regression. In particular, heteroscedasticity is modelled by quantile regression, while the coverage probability of the prediction interval is guaranteed. The method is applicable to all machine learning quantile regression algorithms (Romano et al. 2019; Sesia and Candès 2020).

Conformal prediction in Bayesian settings: Fong and Holmes (2021) combine the benefits of conformal prediction and Bayesian statistical modelling. Regularization may become a feature of the algorithm through using suitable prior distributions, the algorithm becomes more computationally feasible while keeping the coverage properties of conformal prediction.

Other models: Conformal prediction is marginal coverage in the sense that it is not conditioned on the predictor variable. An extension to conditional coverage is introduced by Lei and Wasserman (2014).

5.5.4 Support vector regression

A semiparametric support vector regression variant of quantile regression has been proposed by Shim et al. (2012). Expectile regression with support vector machine regression has been proposed by Farooq and Steinwart (2017), while changing the quantile loss function to an orthogonal one makes support vector regression suitable for noisy data (Yu et al. 2018).

5.5.5 Splines

Quantile regression with splines has been investigated by He and Ng (1999). Non-crossing quantile regression has been explored by Kitahara et al. (2021) and by Xu and Reich (2021) in a Bayesian setting.

5.5.6 Local polynomial fitting

Polynomial quantile regression has been combined with quantile regression trees for probabilistic predictions by Chaudhuri and Loh (2002). Adam and Gijbels (2022) propose expectile regression with local polynomial fitting.

5.5.7 Functional regression

A general framework of regression models for functional data that includes quantile regression, as well as distributional regression, has been proposed by Greven and Scheipl (2017, see also Scheipl et al. 2016). Models are estimated using penalized likelihood or gradient boosting (Brockhaus et al. 2020). The model may include scalar or functional covariates of multiple types (e.g., linear effects, random effects, splines and more).

Regarding functional time series probabilistic forecasting, Hyndman and Shang (2009) propose bootstrap techniques. Moreover, probabilistic univariate time series forecasting has been addressed in a similar context by Shang and Hyndman (2011). Conformal prediction for multivariate functional data has been proposed by Diquigiovanni et al. (2022) and by Kelly and Chen (2022) for multiple functional regression.

5.5.8 Fuzzy logic

Methods for probabilistic forecasting with fuzzy time series has been proposed by Silva et al. (2016, 2020). They are based on specialized techniques that are used in time series forecasting and are not related to those presented in previous sections.

5.5.9 Converting deterministic to probabilistic predictions

The majority of environmental models issue point predictions, also known as deterministic predictions, in the field. To address this, some ideas have emerged to transform them into probabilistic predictions. These methods include the analog

ensemble and the method of dressing, which have been formalized by Yang and van der Meer (2021) for forecasting applications. Given that point predictions are also issued in most machine learning applications, these ideas can be transferred to this field as well.

The analog method (Lorenz 1969) is based on the assumption that a current pattern of a time series is similar to past patterns. Therefore, what was realized in the past is likely to be realized in the future. This method uses strategies to find similar patterns in the past and use them to forecast the future. Applications of analog methods can be found in Alessandrini et al. (2015) and subsequent works. However, it is not always possible to identify similarities between past and current patterns when forecasts are issued by machine learning algorithms (Yang and van der Meer 2021). In this case, the method of dressing is preferred (Roulston and Smith 2003). This method estimates uncertainty by dressing past errors to current point forecasts.

5.5.10 Conditional probability density estimation with kernel smoothing

Kernel smoothing is a prominent non-parametric approach for conditional density estimation. In particular, the conditional probability distribution in a regression setting can be estimated by applying kernel smoothing to the empirical probability density distributions that are components of the conditional probability density function, i.e., (a) the joint distribution of the predictor and dependent variables and (b) the marginal probability distribution of the predictor variables (Hyndman et al. 1996). Kernel smoothing serves two scopes, i.e., (a) to smooth the empirical probability density distributions, while (b) simultaneously weight data points based on their similarity values (Gneiting and Katzfuss 2014). A review of related methods can be found in Rothfuss et al. (2019).

5.5.11 Other models

Other models for probabilistic predictions include the ensemble model output statistics (EMOS) by Gneiting et al. (2005). The method is a variant of linear regression that uses ensemble predictions as predictors and models the conditional mean and variance of a conditional Gaussian distribution. The method resembles distributional regression, but it further exploits properties that are met in weather forecasts when the latter take the form of ensembles. The mean and variance of a possibly non-Gaussian distribution has been modelled also by Nott (2006). The dependent variable is modelled by a double exponential family distribution and predictors are modelled by penalized splines, while inference is performed with a Bayesian framework. Again the proposed model resembles distributional regression. In a similar to GAM modelling context, Fasiolo et al. (2021) propose estimating directly the quantiles of conditional response distribution based on the quantile loss function in a Bayesian setting, thus avoiding to specify the probability distribution of the dependent variable as done in GAMLSS modelling.

Isotonic distributional regression is a case of distributional regression, in which the predictor space is equipped with a spatial order, while the conditional distribution is estimated using suitable loss functions in a way that the partial order is respected. The method exploits the partial order relationships. Compared to distributional regression models (see also previous sections), isotonic distributional regression is automatic after defining the spatial order. Frameworks for isotonic distributional regression are proposed by Henzi et al. (2021a, 2021b).

6 Neural networks and deep learning

Neural networks and deep learning models (i.e., those composed by multiple processing layers; Lecun et al. 2015; Schmidhuber 2015) including their variants (e.g., recurrent neural networks (RNNs), with the special case of long short-term memory (LSTMs), Hochreiter and Schmidhuber 1997, or convolutional neural networks, CNNs) are examined separately in this section due to their increasing significance. Previous reviews on probabilistic predictions have already identified several categories of neural networks and deep learning (see, e.g., Khosravi et al. 2011; Kabir et al. 2018; Ovadia et al. 2019; Abdar et al. 2021). Here, we set the relevant material in the context developed previously in the paper.

6.1 Bayesian methods

Bayesian neural networks work similarly to Bayesian models defined earlier. Priors are assigned to model's parameters (weights on the connections and variance of the error term, Lee 2000) and posterior inference is possible for parameters, resulting in consistent posterior distributions (Lee 2000), while predictive distributions can also be estimated in a similar way. Reviews on Bayesian neural networks can be found in Lampinen and Vehtari (2001) and Titterington (2004). Some variants of Bayesian neural networks include those by Rohekar et al. (2019) who assigned prior distributions that depend on unlabelled predictors and Harva (2007), where the distribution of the dependent variable is a mixture of Gaussian distributions. The latter approach allows modelling the dependent variable with non-Gaussian distributions, thus being more flexible.

Due to heavy computational burden of Bayesian methods, variational inference methodologies have been developed to approximate the posterior distribution of the parameters (Swiatkowski et al. 2019). In particular, a probability distribution of a parametric form is specified and is estimated by minimizing the Kullback–Leibler (KL) between the specified distribution and the true posterior distribution. A likelihood-free variational inference methodology that allows for non-tractable variational densities, resulting in richer structures has been proposed by Tran et al. (2017).

6.2 Monte Carlo simulation-based methods

Since Bayesian techniques are computationally expensive, simulation techniques that can approximate Bayesian outputs may be efficient solutions when probabilistic predictions are of interest. Dropout (Srivastava et al. 2014) is one such technique in which during training, some layer outputs are dropped. Repeating training, one can obtain an ensemble of neural networks. The overall procedure is a regularization method that prevents overfitting. It has been proved (Gal and Ghahramani 2016) that dropout is an approximation of probabilistic deep Gaussian process in a Bayesian context, while it is computationally faster. In particular, predictive uncertainty with dropout can be estimated by calculating the variance of the ensemble.

Batch normalization is another Monte Carlo simulation-based technique that is used for regularization of neural networks. It has been proved that ensembles of neural networks obtained by batch normalization can again approximate uncertainty estimated in a Bayesian context (Teye et al. 2018). Another similar idea by Antorán et al. (2020) is to model the neural network depth as a random variable, resulting in Depth Uncertainty Networks (DUNs). In DUNs, marginalization over depth resembles Bayesian Model Averaging

(BMA; see the explanation on combination methods in Sect. 8.2) over the ensemble of NNs (with each being deeper than previous ones). Another idea based on randomization comes from Mancini et al. (2021), who create ensembles of neural networks by randomizing the architectures (each neural network has different layer specific widths) using a Bernoulli mask. Prediction intervals are created by estimating the variance of the ensemble members.

6.3 Quantile and expectile regression

Probabilistic predictions using scoring functions is a natural choice for neural networks, since an integral part of their structure is exactly the objective function. Quantile regression neural networks have first been proposed by Taylor (2000) with software implementation by Cannon (2011). Xu et al. (2017) and Tagasovska and Lopez-Paz (2019) expand the quantile regression neural network to the case that the loss function is averaged at multiple quantile levels (this technique resembles the quantile scoring rule based on the quantile loss function, see Sect. 3.3.3). Modelling simultaneously at multiple quantile levels also mediates the quantile crossing problem, while also borrows information among various levels, thus producing more accurate predictions.

Related quantile regression neural networks variants include the recurrent neural network by Xie and Wen (2019), in which quantiles at multiple levels are estimated simultaneously. Quantile regression with tensors as predictors has been proposed by Lu et al. (2020). Tensor methods are implemented in deep learning models for improving estimation procedures. Moon et al. (2021) propose learning multiple quantiles with neural networks, by an optimization procedure that imposes an L_1 penalty to the gradient of the loss function. The resulting algorithm addresses the crossing-quantiles problem. In line with Bayesian quantile regression settings, Jantre et al. (2021) develop a similar one for neural networks that inherits merits of Bayesian models (see, e.g., Bayesian statistical modelling and quantile regression in Sect. 4.2.1). Deep learning quantile regression for right censored data for has been proposed by Jia and Jeong (2022). Expectile regression with neural networks by implementing the expectile loss function has been proposed by Jiang et al. (2017). Huber quantile regression with neural networks has been proposed by Tyralis et al. (2023). Quantile regression neural networks and expectile regression neural networks are edge cases of Huber quantile neural networks.

6.4 Distributional regression

Distributional regression with neural networks is natural given the predictive abilities of neural networks. Early developments include Nix and Weigend (1994), in which the outputs of a neural networks are the mean and the variance. Similarly, the algorithm proposed by Cannon (2012) is used to estimate parameters of a target probability distribution using a neural network. In a related context, Lakshminarayanan et al. (2017) train a neural network using a log-likelihood function (recall its relationship with proper scoring rules) to estimate the mean and variance of Gaussian distribution (the mean and variance are outputs of the neural network), while Rasp and Lerch (2018) implement the CRPS for a Gaussian family of distributions. Klein et al. (2021) combined a copula regression model with a neural network that can model a richer family of distributions compared to the Gaussian focused approaches mentioned before. A special case of distributional regression (using the Von Mises distribution) with deep learning for modelling directional (circular) statistics has been proposed by Prokudin et al. (2018). A more general framework for distributional regression

with deep learning has been proposed by Rügamer et al. (2023b, c), while transformation models in deep learning can be found in Kook et al. (2023).

6.5 Other neural networks models

Other variants for probabilistic predictions are mostly based on techniques already presented earlier, but tailored for neural networks. These include the following.

Conformal prediction: Conformal prediction with neural networks has been investigated by Demut and Holeňa (2012) for neural networks and Kompa et al. (2021) for deep learning. The special case of Recurrent Neural Networks (RNN) has been investigated by Stankevičiūte et al. (2021).

Transformation models: Neural network based transformation models have been investigated by Baumann et al. (2021).

Generative adversarial networks: Generative adversarial networks (GANs) are generative models; therefore, their extensions for probabilistic prediction is natural. GANs consist of two deep learning models, specifically of a generative network that generates a sample and a discriminative network that evaluates the sample (i.e., it predicts whether is real or fake in the sense that the sample belongs to the training set). The two networks compete in minimax two-player game regarding their losses (Goodfellow et al. 2014).

GANs are trained by minimizing a statistical divergence. Pacchiardi et al. (2022) proposed training a GAN using scoring rules, aiming to implement them for probabilistic forecasting. Conditional GANs are a formulation of GANS that allows conditioning a model on some data. Those techniques can be interpreted as regression problems in which GANs can provide probabilistic predictions and have been exploited by Koochali et al. (2019) in a probabilistic time series forecasting context and by Koochali et al. (2021) for multivariate probabilistic time series forecasting.

Other variants: Kuleshov et al. (2018) proposed an algorithm that focuses on calibration properties of the regression model, thereby deviating from the principle of maximization of the sharpness of the predictive distribution subject to calibration (see also Sect. 3.3.2; recall also the discussion of the specification of coverage probabilities in Sect. 3.3.3). They implement the algorithm with Bayesian deep learning models. Thiagarajan et al. (2020) proposed using different models for estimating the mean and prediction intervals of the dependent variable, and exploit estimates of prediction intervals for improving mean estimates in a bi-level optimization setting. An empirical assessment shows that their model produces good uncertainty estimates.

Li et al. (2021) propose an alternative solution, in which they partition the range of the dependent variable and use a multi-class classification model to assign probabilities to the predicted dependent variable depending on which bin it falls. The classification probabilities are then easily transformed to distributions. Probabilistic predictions in federated learning have been proposed by Thorgeirsson and Gauterin (2021), who assigned probabilities to the local weights of the model.

7 A representative simple example

To better understand what probabilistic predictions are and how different models can estimate predictive distributions and be compared regarding the accuracy of their predictions we composed an example using simulated data. In particular, we simulated the following linear model:

$$\underline{y} = x_1 + x_2 + \underline{\varepsilon},\ \underline{\varepsilon} \sim \mathrm{N}\left(0, 1^2\right) \tag{33}$$

where $\underline{\varepsilon}$ is standard Gaussian noise. $\underline{x}_1$ and $\underline{x}_2$ were also simulated by a standard Gaussian distribution (recall from Sect. 2.1 that the choice of distribution is irrelevant to the result) and the resulting y was obtained as the sum of Eq. (33). We simulated 100 samples and used 50 samples for training and 50 samples for testing. Quantile regression and a GAMLSS with Gaussian dependent variable were fitted to the training set. Probabilistic predictions of the samples in the test set at multiple quantile levels with both algorithms is presented in Fig. 2.

The simulation was repeated 10 000 times to have an accurate assessment of the performances of both methods using simulations of 100 (see, e.g., the random case in Fig. 2) and 1 000 samples (with 500 samples for training and 500 samples for testing). Coverage probabilities and quantile scores were averaged for the 10 000 cases, for both sizes of samples (100 and 1 000) and are presented in Table 2. Since quantile scores are not scale free, coverage probabilities show the difference in performance of both methods for different sample sizes (but recall that coverage probabilities are not proper scoring rules). Both methods become more accurate as the sample increases. For instance the mean coverage probabilities at level $\alpha = 0.05$ are 0.07344 and 0.05211 for sample sizes 100 and 1 000 for GAMLSS, respectively.

Regarding quantile scores, GAMLSS performs better compared to quantile regression at all quantile levels and all sample sizes. However, the difference between the two methods decreases as the sample size increases. One should expect that GAMLSS would be better because the distribution of the depended variable is well specified (recall that this is a simulation experiment where properties of variables are known a priori), thus giving an advantage to parametric models. As the sample size increases, non-parametric models improve at a higher rate but still remain worse compared to the parametric ones. The situation could be reversed in real data examples where the chance of a misspecified probability distribution of the dependent variable becomes high; therefore, non-parametric methods start to prevail.

8 Combinations of algorithms

Combinations of algorithms, also termed ensemble learning in the machine learning field have been proved to improve over the individual algorithms with respect to multiple aspects, including predictive performance for point predictions. The same can be said for the combination of algorithms that issue probabilistic predictions. That direction of research of probabilistic forecasting has been particularly developed by the forecasting community, following the vast knowledge on the combinations of algorithms that deliver point forecasts. An overall view of the field can be found in Winkler et al. (2019).

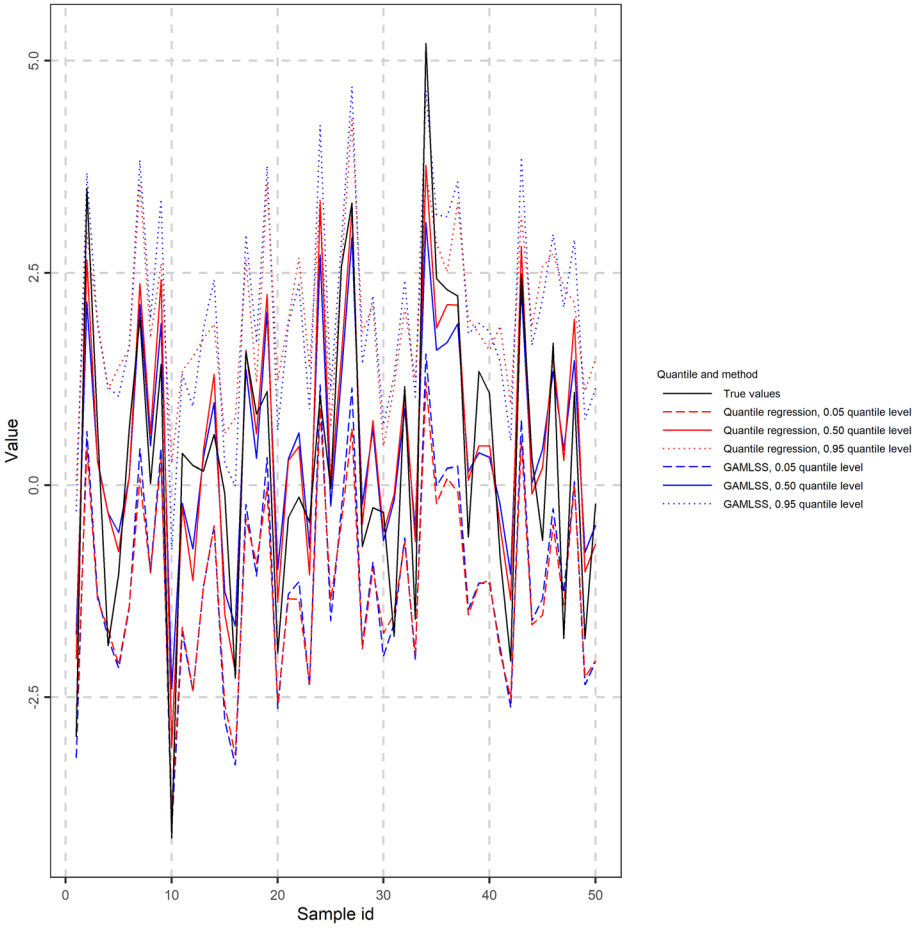

Fig. 2 An example of probabilistic prediction with quantile regression and GAMLSS at quantile levels $\alpha \in \{0.05, 0.50, 0.95\}$ using data simulated by a linear model with Gaussian noise

Table 2 Results of simulation experiments for quantile regression and GAMLSS

Sample size	Model	Coverage probabilities			Quantile scores		
		$\alpha = 0.05$	$\alpha = 0.50$	$\alpha = 0.95$	$\alpha = 0.05$	$\alpha = 0.50$	$\alpha = 0.95$
100	Quantile regression	0.07921	0.49943	0.92112	0.12044	0.41847	0.11988
100	GAMLSS	0.07344	0.49890	0.92677	0.11460	0.41317	0.11426
1 000	Quantile regression	0.05281	0.50034	0.94728	0.10456	0.40095	0.10455
1 000	GAMLSS	0.05211	0.50026	0.94793	0.10391	0.40027	0.10389

8.1 Weighted averaging of algorithms

The simplest combinations are those that are based on weighted averaging of predictions.

Now there are some different directions that can be followed. In the first crossroad, there is the dilemma of combining distributions or functionals of distributions (e.g., quantiles), with the former being met more frequently. After deciding on what to combine, a method to estimate the weights of the combination is needed. Combinations of algorithms may include, for instance, the weighted linear averaging (the "*linear opinion pool*") of cumulative distributions (Eq. (34)) or densities (Eq. (35)), or the multiplicative combination (the "*logarithmic opinion pool*") in Eq. (36); see, e.g., the overview by Genest and Zidek (1986).

$$F = \sum_{i=1}^{K} w_i F_i \tag{34}$$

$$f = \sum_{i=1}^{K} w_i f_i \tag{35}$$

$$f = \prod_{i=1}^{K} f_i^{w_i} \Big/ \int \prod_{i=1}^{K} f_i^{w_i} \mathrm{d}\mu \tag{36}$$

In the right-above equations, K is the number of algorithms to be combined, and F_i and f_i are the cumulative distributions functions (CDFs) and probability density functions (PDFs) predicted by algorithm i, respectively. Several other combination strategies also exist, including the centered linear pool (Knüppel and Krüger 2022). The expansion to the combination of quantiles or other functionals (e.g., expectiles) is straightforward.

The equally weighted linear pool (simple averaging) is a combination that is hard to beat in practice (Lichtendahl et al. 2013), with multiple attempts being made and being more successful as the data size increases (see, e.g., the comparison by Clements and Harvey 2011 and the theoretical comparison by Gneiting and Ranjan 2013).

Regarding the choice of averaging distributions or quantiles when examining equally weighted linear pools, Lichtendahl et al. (2013) support the latter option, which is shown that it assigns sharper forecasts. Furthermore, Lichtendahl et al. (2013) have shown that the equal weight averaging performs no worse compared to the average of the scores of the individual algorithms. This finding is called "*harnessing the wisdom of the crowd*".

Random forests and boosting are ensemble learning algorithms. A single type of base learner is used (usually decision trees) by both algorithms. In Sects. 5.3 and 5.4, we presented how random forests and boosting can assign probabilistic predictions. Stacking needs diverse types of algorithms and resembles the culture of forecasters in combining algorithms. We shall talk about stacking in Sect. 8.3, while earlier in Sect. 8.2 we shall talk Bayesian Model Averaging (BMA), which seems to be another standard practice of forecasters in combining probabilistic forecasts.

Combining distributions: Minimizing a loss function with respect to the weights of the implemented opinion pool is a usual method for estimating optimal weights. Hall and Mitchell (2007) proposed estimating the weights of the linear pool by minimizing a distance, measured by the Kullback–Leibler information criterion (which corresponds to the log-score), while Hora and Kardeş (2015) proposed estimation by minimizing the quadratic scoring rule. Opschoor et al. (2017) and Berrisch and Ziel (2021) estimated weights by minimizing the censored likelihood scoring rule as well as the CRPS. Ranjan and Gneiting (2010) recommended a beta-transform of the linear opinion pool and

they estimated the parameters of the beta-transform and the weights by maximizing the likelihood of the model (recall the relationship between maximum likelihood and the logarithmic score in Sect. 3.4). The model is shown to improve over the linear opinion pool with regards to calibration and sharpness.

A more flexible option to allow estimated weights depending on the forecasted value, thus not being constant has been proposed by Kapetanios et al. (2015). Bayesian modelling of beta-transformed opinion pools has been proposed by Casarin et al. (2016) and Bassetti et al. (2018) resulting in better performances. Previous combinations refer to multiple predictions for a single point. Ray et al. (2017) proposed combining probabilistic time series forecasts at multiple horizons using copulas.

Combining quantiles: Estimation of quantile forecasts by minimizing a quantile loss function with respect to the weights has been proposed by Taylor and Bunn (1998), i.e., to apply quantile regression to linear opinion pool. Since linear in parameters quantile regression allows for a constant term in addition to the weights of the covariates, that constant term is also possible to be added in the linear pool. Taylor and Bunn (1998) have shown that, when quantile predictions are unbiased, the constant should be omitted and the weights should sum to unity. In Shan and Yang (2009), the weights are estimated by a ratio of multiplicative quantile losses (after being multiplied with a tuning parameter and transformed with an exponential function) divided by the loss of all algorithms.

Methods for the combination of prediction intervals have been investigated by Gaba et al. (2017) that may also apply to the combination of quantile forecasts and vice versa, since prediction intervals are defined by their upper and lower predictive quantiles. Besides the simple average, such methods include the median, the envelope of the interval, exterior as well as interior trimming and more.

8.2 Bayesian model averaging

Besides averaging algorithms, a natural alternative for combining models is in Bayesian settings, in particular Bayesian Model Averaging (BMA) (for a review on BMA, see Hoeting et al. 1999 and Kaplan 2021). Assuming that models M_i, $i = 1, \ldots, K$ are combined and z is the (future) variable of interest (recall the notation from Sect. 2.1), then the predictive distribution is given by

$$p(z|y) = \sum_{i=1}^{K} p(z|M_i, y)p(M_i|y) \tag{37}$$

where $p(M_i|y)$ is given by.

$$p(M_i|y) = (p(y|M_i)p(M_i))/(\sum_{i=1}^{K} p(y|M_i)p(M_i) \tag{38}$$

and $p(y|M_i)$ is given by.

$$p(y|M_i) = \int p(y|\theta_i, M_i)p(\theta_i|M_i)\mathrm{d}\theta_i \tag{39}$$

in the usual Bayesian way.

Applications of BMA include improvements of forecasts ensembles that tend to be underdispersive (Raftery et al. 2005). Several options for assigning distributions to the ensemble members are possible, depending on the problem at hand. For instance, Baran (2014) implemented truncated normal distributions. Other applications include imputation

of missing data (Kaplan and Yavuz 2020). BMA has also been used as an inherent component of machine learning models. For instance, it was applied to BART (Hernández et al. 2018; see Sect. 5.5.1) to obtain an algorithm that performs better when the number of predictors is large.

8.3 Stacked generalization

Stacked generalization (also called stacking; Wolpert 1992) is an ensemble learning algorithm that gains strength when combining machine learning algorithms with diverse properties. In contrast to combinations algorithms presented in Sect. 8.1, in which weights are estimated based on in-sample performance, stacking weights are estimated by out-of-sample performances. Implemented losses can be consistent scoring functions or proper scoring rules as those presented in Sect. 3.

Yao et al. (2018) favour the use of stacking over BMA for probabilistic predictions in cases that the set of combined algorithms does not include the true data generating model. The combiner algorithm could be linear (e.g., the weighted average in Sect. 8.1), but more flexible algorithms (e.g., quantile regression forests) could also be used for increasing performance.

9 Special cases

Existing machine learning algorithms are adapted for specific applications for which some prior knowledge exists regarding properties of the problem under investigation. For instance time series forecasting problems are characterized by temporal dependence structures and spatial prediction problems are characterized by spatial dependence structures. Exploiting information from those dependence structures may result in improved predictive performances. Specialized algorithms for temporal, spatial and spatio-temporal models are presented in Sects. 9.1–9.3.

Other specialized applications of probabilistic predictions that attract interest due to being especially relevant in scientific fields are also presented in the following. Those include, for example, predictions of extremes and uncertainty of measurements (see Sects. 9.4 and 9.5, respectively). We conclude with applications of probabilistic predictions in various scientific fields (see Sect. 9.6).

9.1 Temporal models

A treatment of Gaussian processes for time series forecasting including the case of probabilistic forecasts can be found in Roberts et al. (2013). Modelling time series with an algorithm that borrows ideas from GAMs can be found in Taylor and Letham (2018). The proposed algorithm models various time series components, including seasonality, trends and effects of holidays, while it also assigns probabilistic forecasts.

Other models are mostly deep learning based with special features, following a recent explosion of deep learning use in forecasting competitions, although tree-based methods were also successful in those competitions (Januschowski et al. 2021). In fact, most deep learning algorithms are based on RNN architectures (Hewamalage et al. 2021), albeit exceptions also exist. Deep learning algorithms for probabilistic forecasting include:

Hybrid models: Those models integrate two or more algorithms and exploit properties from both algorithms. For instance, Khosravi et al. (2013) use a neural network to predict the mean of a time series and, subsequently, the predictions are transformed to probabilistic by imposing a GARCH model on top of the neural network to model the variance. Gasthaus et al. (2020) propose modelling the distribution of the response variable using spline quantile functions on top of a RNN, instead of modelling quantiles separately at each level. Smyl (2020) proposed a model that combines RNNs and exponential smoothing. The role of the latter algorithm in the model was to deseasonalize and normalize the time series.

Quantile regression models: Quantile autoregression neural networks that include neural networks combined with quantile autoregression have been proposed by Xu et al. (2016). A quantile regression neural network for mixed sampling frequency data has been proposed by Xu et al. (2021).

Monte Carlo-based methods: Implementation of dropout in LSTM (a type of RNN that is particularly resistant to optimization misbehaviours) for probabilistic forecasting has been proposed by Serpell et al. (2019). Bootstrap-based predictions intervals were forecasted by a neural network by Mathonsi and Van Zyl (2020).

Distributional regression: The method by Hu et al. (2020), in which the parameters of a variable bounded in [0, 1] is estimated by neural networks can be said to belong to the class of distributional regression. Variables with larger support can be transformed to ones bounded in [0, 1].

Modelling multiple time series simultaneously: Modelling multiple time series with a single model can exploit additional information compared to the case that each series is modelled separately and leads to large improvements regarding predictive performance (Sen et al. 2019). Examples include Chen et al. (2020), who implement a CNN architecture and allow for quantile and distributional regression, and Salinas et al. (2020), who implement an autoregressive component, while their model is of the distributional regression type.

Software: Software that unifies multiple models for probabilistic time series forecasting include Gluon Time Series Toolkit (GluonTS) by Alexandrov et al. (2020).

9.2 Spatial models

As already stated previously, dependence structures in spatial problems can be exploited to improve predictions similarly to time series models. However, compared to the usual case in time series modelling, an additional obstacle in spatial modelling is that measurements are placed in irregular locations. Machine learning models are also used for spatial predictions, among which Gaussian process regression holds a prominent place due to tradition in the field. Pure machine learning models are also applicable, although information from spatial dependencies is not exploited in a straightforward way (Hengl et al. 2018).

Gaussian process regression: A Bayesian statistical model for Gaussian process modelling has been developed by Banerjee et al. (2008). The model is particularly suited for big datasets, while also remedying relevant weaknesses that are inherent characteristics of Bayesian models. Further developments to address issues of computational complexity with computationally efficient modelling have been approximate Bayesian inference by Eidsvik et al. (2012), convenient prior distributions (Duan et al. 2017) and nearest-neighbour Gaussian processes (Zhang et al. 2019). Inserting Gaussian processes to a multi-layer model is another approach that exploits advances in deep learning (Zammit-Mangion et al. 2021).

Markov random fields: Markov random fields can be used to approximate spatial processes and are more appropriate for measurements in a regular grid (Banerjee et al. 2008). For a treatment of Gaussian Markov random fields, see Rue and Held (2005) and MacNab (2018). A GAMLSS approach to Gaussian Markov random fields has been proposed by De Bastiani et al. (2018).

Quantile regression: Quantile regression has been extended to the spatial case by Hallin et al. (2009). The spatial dependence has been modelled by copulas in a Bayesian statistical-based setting (Chen and Tokdar 2021).

Additive models: Bayesian modelling for additive models that may also include Markov random fields and spatial effects has been examined by Fahrmeir and Kneib (2009).

Bayesian model averaging: Bayesian modelling averaging adapted for spatial settings (termed geostatistical model averaging) has been proposed by Kleiber et al. (2011).

Deep learning: Sidén and Lindsten (2020) showed a connection between Gaussian Markov random fields and CNNs that allows generalization to multi-layer architectures (termed deep Gaussian Markov random fields).

9.3 Spatio-temporal models

Modelling of temporal and spatial data (Hefley et al. 2017) poses new challenges, while new more complex models are needed to address related problems. Those models have been particularly developed by the community of the spatial statistics field, and are mostly based on Bayesian statistics, with characteristic examples being available in Katzfuss and Cressie (2012) and Finley et al. (2015). Gaussian processes have also been used (Wang and Gelfand 2014). Expectile-based models (Spiegel et al. 2020) and deep learning (Wu et al. 2021) have also become options to model spatio-temporal data.

9.4 Extremes

Predicting events that happen infrequently is a major challenge for machine learning. Such events can be various disasters (e.g., floods) and unexpected events (e.g., extremely high demand for electricity that cannot be supported by the grid). When prediction of a variable is probabilistic and the interest is on modelling extremes, then the tails of the distribution or extremely low or high quantiles of the distribution should be modelled. Distributions that are suitable for such variables are usually heavy-tailed, i.e., a larger part of the mass is located at the tails of the distribution compared to light-tailed distributions (e.g., the Gaussian distribution). Estimation of the parameters in distributional regression approaches may be unstable at such cases, while quantile regression (or consistent scoring functions-based methods) may not suit for such problems, since conventional large-sample theory does not apply sufficiently far in the tails (Chernozhukov 2005).

In the following, we will discuss issues related to scoring functions for extremes and improvements on quantile regression for modelling extremes:

Quantile regression: Large sample properties of extremal quantile regression have been studied by Chernozhukov (2005). Wang and Li (2013) and Wang et al. (2012) proposed extrapolating intermediate quantiles, predicted with quantile regression, far in the tails using extreme value theory. Firpo et al. (2021) applied a generalized method of moments (GMM) estimation of quantile regression coefficients by using flexible parametric restrictions that allow to extrapolate intermediate quantiles.

Expectile regression: Extrapolation of intermediate expectiles far in the tails has been proposed by Daouia et al. (2018), while functional estimation has been proposed by Girard et al. (2022a). Furthermore, a non-parametric method for predicting extreme expectiles has been developed by Girard et al. (2022b) based on their distributional relationship with quantiles. Joint inference for several extreme expectiles has been investigated by Padoan and Stupfler (2022).

Scoring extremes: Diks et al. (2011) proposed using as scoring rules, a conditional likelihood, given that the event lies in the region of interest (in the tail of the distribution) or a censored likelihood, with censoring of events outside the region. Those scoring rules are more appropriate when the interest in on predicting the distribution at the specified region. Closed form expressions of the CRPS for the (heavy-tailed) generalized extreme value distribution (GEV) and the generalized Pareto distribution (GPD) were developed by Friederichs and Thorarinsdottir (2012). Weighted scoring rules with emphasis on the tails have been proposed by Lerch et al. (2017). The exceedance probability scoring rule, which is suitable to assess exceedance probabilities, was developed by Juutilainen et al. (2012), while Taggart (2022a) proposed consistent scoring functions for comparing point predictions in the region of tail of the distribution. Non-parametric scoring methods that are based on a cross-validation setting were also proposed by Gandy et al. (2022). However, expected scores are not appropriate to distinguish tail properties (Brehmer and Strokorb 2019).

Other methods: Oesting et al. (2017) proposed Brown-Resnick stochastic processes for forecasting extreme events. Brown-Resnick processes are max-stable; thus, they can model heavy tails, while they are also particularly suitable to model spatial random fields with max-stability properties.

Deep learning: Deep learning (in particular CNNs) has been used to estimate the type of dependence (specifically, to guess between asymptotic dependence or independence) of spatial extremes (Ahmed et al. 2021).

9.5 Uncertainty in measurements (observational errors)

Measurement errors are common in many fields (Hariri et al. 2019). For instance, measured quantities of rainfall over an area tend to be inexact since stations cannot cover the full region. Regression models are tailored to predict based on data with errors; thus, they do not report the real values. For this reason, adaptations of regression models are needed to account for these errors. Methods that account for several types of measurement errors have been developed when probabilistic predictions are needed. Some representative methods will be presented in the following.

Wei and Carroll (2009) developed a quantile regression method that corrects measurements errors in the predictors by adapting the quantile-loss based minimization problem, introducing a probability distribution to model the error. The case of measurement errors of the dependent variable has been examined by Ferro (2017) and Bessac and Naveau (2021), who adapted proper scoring rules to address those errors (see also Brehmer and Gneiting 2020, for a related example on proper scoring rules).

9.6 Some representative applications

The requirement for delivering probabilistic predictions has been recognized in several scientific fields. Consequently, review articles on aspects of methods for probabilistic predictions are met frequently. In the following, we discuss them.

9.6.1 Economics and finance

In economics and finance, Duran (2008) discusses methods for probabilistic forecasting of sales. Deep learning for financial time series forecasting has been surveyed by Sezer et al. (2020). Although, most applications regard point forecasts, the paper is a good starting point for probabilistic time series forecasting given the straightforward extensions to probabilistic deep learning, as already discussed earlier here. Finally, combination of algorithms in economics has been reviewed by Steel (2020).

9.6.2 Energy research

Applications of probabilistic predictions are frequent. They are summarized by papers that are discussed in the following.

Renewable energy: Probabilistic wind power generation forecasting has been reviewed by Zhang et al. (2014). Yang et al. (2022) reviewed solar forecasting including probabilistic methods, while van der Meer et al. (2018) surveyed probabilistic photovoltaic power production forecasting. Post-processing in solar forecasting has been reviewed by Yang and van der Meer (2021). Post-processing refers to correcting point forecasts (usually assigned by numerical weather models) and transforming them to probabilistic. Deep learning for wind and solar energy forecasting has been surveyed by Alkhayat and Mehmood (2021).

Electricity price: Electricity price forecasting, including methods for probabilistic forecasting, has been reviewed by Weron (2014). Nowotarski and Weron (2018) and Ziel and Steinert (2018) focused on methods for probabilistic electricity price forecasting.

Electric load and electricity consumption: Probabilistic electric load forecasting has been surveyed by Hong and Fan (2016). Van der Meer et al. (2018) surveyed probabilistic electricity consumption forecasting.

Smart energy systems: Probabilistic forecasting in smart grids has been surveyed by Khajeh and Laaksonen (2022) and Ahmad et al. (2022).

Other topics: Gensler et al. (2018) focus on directives for selecting metrics for probabilistic forecasting in renewable energy. The class of boosting algorithms that are especially relevant for probabilistic predictions in energy has been surveyed by Tyralis and Papacharalampous (2021).

9.6.3 Environmental and earth and planetary sciences

Applications of probabilistic predictions and forecasts are frequent in environmental, as well as in earth and planetary sciences, with machine learning playing an important role in those applications (Haupt et al. 2022). Beyond probabilistic predictions, related topics of interest include the prediction of extremes (Ghil et al. 2011; Sillmann et al. 2017; Huser 2021) and post-processing methods, with the latter topic appearing due to

the need of integrating weather forecasting models, which are especially relevant to the field. We discuss them in the following.

Post-processing methods: Post-processing methods for weather and hydrological forecasting are reviewed by Li et al. (2017) and Vannitsem et al. (2021).

Other topics: Methods for probabilistic predictions in geology are surveyed by Albarello and D'Amico (2015). Probabilistic forecasting with machine learning in hydrology is reviewed by Papacharalampous and Tyralis (2022). Metrics for probabilistic predictions are surveyed by Huang and Zhao (2022) in hydroclimatology. The class of random forests algorithms that can also assign probabilistic predictions has been surveyed in hydrology by Tyralis et al. (2019b).

9.6.4 Big data comparisons

Comparison of multiple machine learning methods using big datasets is essential to understand their performance and allows to understand their properties. Such comparisons exist in the literature, although they are rare compared to comparisons of machine learning methods for point predictions (for the latter, see, e.g., the survey by Fernández-Delgado et al. 2019). Such studies include the comparison of multiple algorithms in environmental data (Papacharalampous et al. 2019; Tyralis et al. 2019a), computer science problems (Torossian et al. 2020), data competitions (Grushka-Cockayne and Jose 2020) and spatial problems (Fouedjio and Klump 2019). Such data comparisons could be facilitated in the future by the implementation of specialized software (e.g., Ghosh et al. 2022).

10 Summary and future outlook

10.1 Setting a probabilistic prediction problem—technical considerations

Setting a probabilistic prediction problem requires understanding of all nuances in the field. That is possible by knowing the evolution in the field from the simplest statistical algorithms to more complex ones. As it was shown previously in this review paper, the foundation of the field lies on simple concepts developed at different time periods.

The first type of foundational concepts originate from the theory of Bayesian statistical modelling. Within this framework, parametric models can estimate the predictive distribution of the dependent variable, given that the model is correctly specified. Another foundational area lies in linear regression modelling. Linear regression models are among the simpler models and can be transformed to Bayesian statistical models to subsequently predict probabilities for future dependent variables. The third foundational area is consisted of scoring rules for training—calibrating statistical models. It allows: (a) fitting parametric models, thereby avoiding Bayesian-based simulations; and (b) fitting non-parametric models, thereby bypassing the problem of the possible misspecification of probability distributions. Based on the three above-outlined foundational areas, as well as on the theory of machine learning models, it is possible to build new models tailored to the problem at hand. Specific problems are already discussed in previous sections of this review paper, while a synthesis of seemingly unrelated methodologies follows. Those methodologies can serve as components of new algorithms for probabilistic predictions in ways that are clarified later in this section.

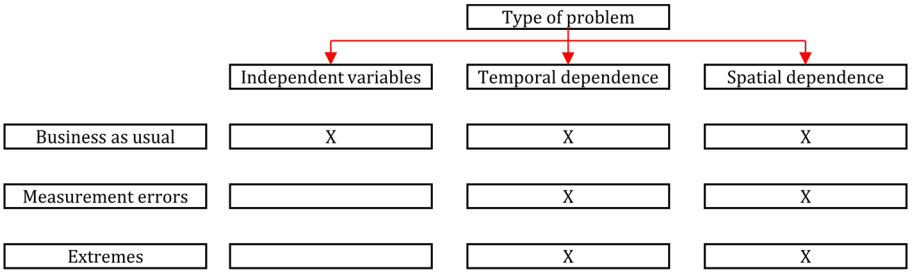

Fig. 3 Frequent types of probabilistic prediction problems and cases of special interest met at each of them

There are three main types of practical problems depending on the readily available information (see Fig. 3). In the first one, the response variables are independent given the predictor variables. Most machine learning models are suited for this case. The case of temporal dependence (appearing when modelling time series) is well dealt by the time series models, as well as by specialized deep learning models (e.g., LSTMs) that focus on the exploitation of temporal dependence information. Obviously, the usual machine leaning models can also be applied to these cases, with occasionally great performance. A reason for their possible great performance in modelling time series data, despite their inability to exploit the information of temporal dependence, could be that their additional complexity and flexibility compensates for the lost information. Problems with spatial data, in which spatial dependence information might be exploited are usually addressed with spatial statistical models (e.g., kriging) and Gaussian processes. The latter models are preferred compared to other machine learning models in the field of spatial modelling perhaps due to reasons of tradition, although other machine learning models may also be applicable. Regarding the predictive performance in modelling in spatial settings using usual machine learning algorithms, similar arguments with the case of modelling time series using these same algorithms apply.

"Business as usual" modelling is frequently met in all the three aforementioned types of practical problems. On the contrary, modelling of measurement errors and modelling of extremes are mostly required in time series and spatial settings. Although, the latter two modelling cases might also be observed in practical problems with independent variables, these occurrences seem to be relatively rare. Perhaps the main reason behind this fact, for the case of measurement error models, is that some types of observations are characterized by minimal error. This might indeed hold, for instance, in house pricing modelling, where modelling of independent dependent variables is effective. Modelling of extremes is met in environmental and finance applications, in which temporal or spatial dependence seems to be the norm.

Beyond the type of problem, the type of prediction is also of interest. For instance one might be interested in a specified functional of the predictive distributions. That might be, for example, the case in which one is interested on the probability that temperature exceeds a pre-specified quantity. The full predictive distribution of the dependent variable will be more informative, although its estimation might be harder. Those two problem definitions are presented in Fig. 4.

When one is interested in predicting a functional, he/she should apply a consistent scoring function for this functional to fit a related machine learning model. For instance, a quick way to predict a quantile at a specific level is to fit a linear model using the quantile loss function. If one is interested in increased performance, then many algorithms may be

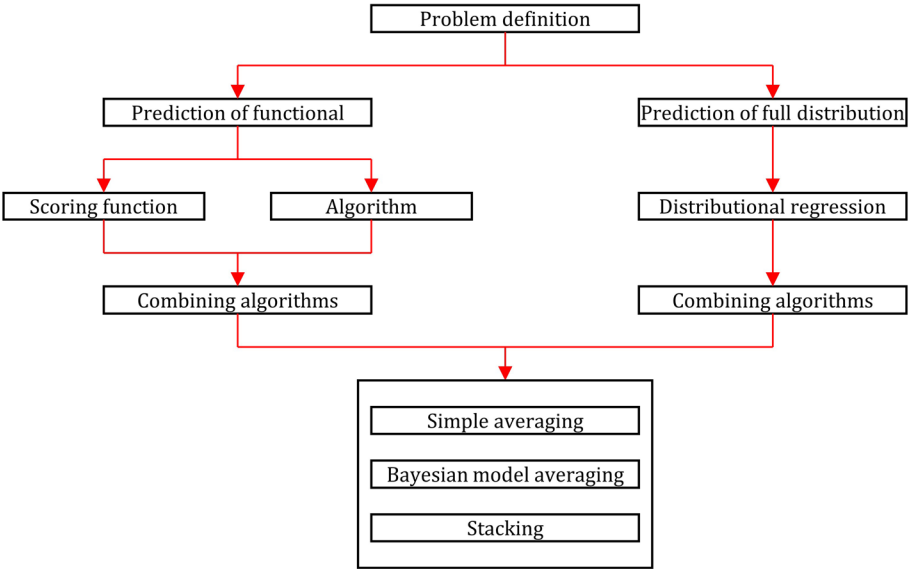

Fig. 4 Problem definition based on type of prediction and related ways for handling such problems

fitted and then combined. The primary combiners are simple averaging, Bayesian model averaging and stacking, with each method having its merits and deficiencies. Similarly, distributional regression is the primary way to obtain the full predictive distribution, while fitting can be done with proper scoring rules. Again, the combination of algorithms is possible for obtaining improved predictive performance. Obviously, by estimating the full predictive distribution, functionals can be made readily available, while by estimating multiple functionals (e.g., quantiles at multiple levels) one can also get the predictive distribution at a high resolution. Both approaches have their merits, which are discussed in Sect. 10.2.

Figure 5 is representative on how elements from the three foundational areas can be combined to have a new algorithm that can be suited for a specific task. For instance, one might be interested in estimating the full predictive distribution using boosting. In this case, the most suitable algorithm might be the distributional boosting one described in Sect. 5.4. This algorithm is based on the optimization of a proper scoring rule applied to a machine

	Simulation – based methods	Point prediction methods	Distributional regression
Linear models	X (4.1)	X (4.2)	X (5.2)
Gaussian process regression	X (5.1)		
Random forests	X (5.1)		X (5.1)
Boosting algorithms		X (5.4)	X (5.4)
Neural networks	X(6.2)	X (6.3)	X (6.4)
Other models		X (5.5)	X (5.5)

Fig. 5 Primary algorithms for probabilistic prediction and a classification of them. The numbering in parentheses refers to the section, in which the description of the method can be found

learning algorithm. Another example is that of linear Bayesian statistical modelling, which is founded on the application of Bayesian modelling to a statistical learning algorithm (specifically, to the linear regression model), as described in Sect. 4.1. The expansion of this approach for dealing with the case of time series models leads to the various Bayesian linear time series models (see Sect. 4.3.1). In general, random forests are algorithms that are based on simulation frameworks. Boosting algorithms are based on the optimization of appropriate loss functions, while neural networks seem to be more universal in the sense that implementations are expanded in all the categories. Obviously, exceptions to this general pattern exist, as it is shown in Fig. 5.

Questions about which scoring function or scoring rule to use to assess a point or probabilistic prediction often arise. A consistent scoring function should be used to assess predictions of a specified elicitable functional. However, there may exist a family of consistent scoring functions for the same functional. In this case, it is desirable to disclose the scoring function to the modeller, as consistent scoring functions for the same elicitable functional may provide different assessments. Under three assumptions, the choice of the consistent scoring function will not affect the assessment (Patton 2020):

(a) The information sets of the modellers are nested and do not lead to optimal predictions that are identical.
(b) The predictions are based on models that are perfectly estimated.
(c) The models are correctly specified for the functional of interest.

Unfortunately, these assumptions are rarely met in practice. Even if the expected score is minimized, the choice of the scoring rule will affect the assessment results in finite samples, unless the three assumptions mentioned above are met (Patton 2020). In summary, the choice of the scoring function or scoring rule is an important consideration when assessing predictions. The scoring function should be disclosed to the modeller, and the assumptions underlying the scoring function should be carefully considered.

10.2 Differences between point prediction, distributional regression and Bayesian methods

Knowing the merits of each modelling approach is essential for deciding whether to use models based on consistent scoring functions or models based on distributional regression, when probabilistic predictions are required. Our considerations follow Waldmann (2018), who supports the use of quantile regression in specific situations. They also follow Rigby et al. (2013), who compare quantile regression with GAMLSS.

When one is interested in a specific elicitable functional, then he/she can apply a machine learning algorithm trained with a strictly consistent scoring function for this functional. That is, e.g., the case in probabilistic forecasting competitions where a usually scaled quantile scoring function is disclosed to the competitors. In such competitions, machine learning quantile regression algorithms (e.g., deep learning (Sect. 6.3), boosting (Sect. 5.4), quantile regression forests (Sect. 5.3), support vector regression (Sect. 5.5.4)) are usually among the most successful ones. Predicting the functional is also possible using distributional regression. In this case, the predictive distribution is estimated and then it is straightforward to compute the predictive functional of interest. One can also use Bayesian statistical modelling.

A possible problem arising when using either distributional regression or Bayesian statistical modelling, is the misspecification of probability distributions. Indeed, in most practical situations, the probability distribution of the dependent variable is not known, while setting a probability distribution imposes strong assumptions. In some fields, such as the environmental sciences, some large-scale studies give guidance on the type of the most suitable distribution to model specific variables, although it is still possible that in many sub-cases of the problem at hand, this distribution may not be useful. One could overcome this limitation by using distributions that are more flexible in the sense of modelling more aspects of the dependent variable using many parameters. On the other hand, point prediction algorithms do not encounter this problem, since they are non-parametric and, thus, the additional assumption of probability distribution specification does not limit their flexibility.

On the other hand, the flexibility of point prediction algorithms may be prohibitive in cases that the available data are limited. In those cases, the algorithm may fail to converge rapidly. Therefore, some additional prior knowledge on the probability distribution may be exploited at the cost of reduced flexibility of the algorithm.

Another problem arising mostly in cases of simulation (e.g., Bayesian statistical) models, is the computational and storage cost. In such cases, one should simulate the full distribution that may be computationally prohibitive while also keeping the full sample for future use. That may be impractical if a single functional is of interest, in which case point prediction methods should be preferred. When the size of the data is small, then Bayesian modelling may become more appealing with respect to this fact. Progress in approximate Bayesian computation may also improve the practicality of Bayesian statistical models, while in some cases, part of the calculations is based on explicit formulas that can effectively reduce the need for additional computations. Unfortunately, in the latter case some additional assumptions may also be needed. A relevant example is the use of reference priors in Bayesian statistical modelling.

The size of the data is also important when one estimates high or extreme quantiles. In small data size cases, point prediction methods are not efficient for predicting such quantities, as it has been proven by large-sample theory. Distributional regression approaches might be preferable, although point-prediction algorithms can be adapted to deal with such cases. On the other hand, big data may also allow a suitable specification of the predictive distribution.

When the full predictive distribution of the dependent variable is of interest, Bayesian modelling and distributional regression are natural choices. However, it also possible to estimate functionals, such as quantiles at multiple levels, thereby estimating a substitute of the predictive distribution. That approach eliminates some of the advantages of point prediction methods regarding the speed of calculation, while quantile crossing becomes a possibility. Some advances remedy those undesirable properties, including for instance the simultaneous estimation of multiple quantiles using deep learning models.

In time series forecasting, the application of Bayesian modelling techniques is even harder, due to the larger number of parameters that are needed to be modelled and also due to the imposition of dependence structure. However, again some time series may be short. In this latter case, Bayesian modelling might be benefitting.

Scoring functions for all tasks have not been discovered yet, while the dependent variable may be characterized by some peculiarities, with intermittency consisting a characteristic example of the latter. In such cases, distributional regression might be more suitable. Indeed, for the example of intermittency, modelling can be effective when using appropriate probability distributions (Ziel 2021).

Tradition in point prediction dictates that different models are assessed in a test (out-of-sample) set and the modeller selects the model with the better performance. A scoring function that is consistent for the functional of interest is used to rank the models. That cross-validation-based tradition is rarely met in Bayesian statistical modelling and to a lesser extent in distributional regression. That is reasonable, since Bayesian statistical modelling mostly focuses on parameters inference. Comparison of different models is based on prediction coverages and widths of predictions intervals, although those scoring rules are not proper.

10.3 Final remarks

Several challenges for advancing the field of probabilistic prediction are related to tasks that existing algorithms are not able to complete. These tasks include but are not limited to:

Algorithms tailored to client's requirements: Existing algorithms are mostly trained with prespecified scoring functions or scoring rules. Those scoring rules may not meet users' needs. For instance, estimation of quantiles may not be adequate for quantifying the impact of exceedance, while the magnitude of the exceedance may also be of interest. New scoring functions may be properized to be made consistent for functionals. Optimization of algorithms using proper scoring rules beyond local ones is also a strategy for obtaining improved probabilistic predictions.

Temporal and spatial predictions: Although models for time series forecasting and spatial prediction have been developed, several topics have place for improvements. Those include, for example, the prediction of multiple points simultaneously. Related scoring rules for adapting algorithms exist (e.g., the energy score); however, these are rarely met in practice. The task is more challenging compared to the case of predicting the mean of multiple variables simultaneously. For instance, a problem that has been rarely addressed is that of predicting multiple quantiles, possibly at different levels, simultaneously.

Prediction of extreme quantiles: Predicting extreme quantiles of probability distributions is another challenge and hot topic of current research. The scoring functions for the task are not satisfying so far. Distributional regression may be an option, but is accompanied with several disadvantages (which were already mentioned). The current state of research seems to be heading to a direction that intermediate functionals are predicted and then they are extrapolated using extreme value theory. Other scoring functions that exploit information beyond the frequency over specified levels seem to also be a trending approach.

Observational errors: Modelling of observational errors seems to witness slow progress. Perhaps this is due to the lack of suitable large datasets. Progress seems to be heading towards the direction of developing new scoring functions, as well as towards obtaining progress in Bayesian statistical modelling.

Large-scale comparisons: While large scale comparisons are regularly met for the case of machine learning algorithms for point predictions and forecasts (with these predictions and forecasts being related to means or medians of the dependent variable), that does not seem to be the case for probabilistic predictions. Such comparisons may reveal properties of algorithms based on empirical evidence. When the task is forecasting, time series are of varying magnitudes; therefore, skill scores might be more suitable for comparing competing algorithms. However, such scores may not be proper. Development of statistical tests of the significance of the difference in the performance of different methods still witnesses some small progress. Combining multiple algorithms might also be a way for

improving predictive performance and much is left to be desired towards this direction in the respective field.

Other topics: Other challenges are related to the proper scoring of prediction intervals. Historically, prediction intervals seem to be among the first functionals of interest that were predicted in time series forecasting. The usual strategy for comparing algorithms was based on a metric of calibration (specifically, on coverage probabilities) and on a metric of sharpness (specifically, on widths of prediction intervals). Although they can be informative, such comparisons are not consistent to the task of predicting the functional. Related scoring rules should be used; however, these scoring rules mostly refer to fixed quantile levels. Other approaches for scoring predictions of intervals with varying quantile levels should be developed. Another challenge is related to the prediction of dependent variables with peculiar properties, such as intermittency, multiple modes, heavy tails and more. New scoring rules may help for calibrating machine algorithms to achieve the specific task.

10.4 Final remarks

Although, the various applications of probabilistic predictions are increasing, the vast majority of the machine learning applications refer to point predictions of the mean functional of the distribution of the dependent variable. Due to reasons of tradition, as well as for making the communication of application products easier for their users, the current situation (which is in favour of point predictions) is not expected to change. However, we anticipate that the applications of probabilistic predictions using machine learning algorithms will increase their share, both in the academia and in the industry. We also anticipate that the procedure of users' decisions will adapt to the increased information provided by probabilistic predictions. Further progress in developing new algorithms is also expected to meet the increasing needs.

Acknowledgements The authors are deeply grateful to the Editor for his careful handling of the review process and to the Reviewers for their insightful and constructive comments, which have resulted in a significantly improved article. The Reviewers' suggestions have helped us to expand the scope of the article, clarify the arguments, and improve the overall quality of the writing.

Author contributions HT and GP contributed equally to this work.

Funding Open access funding provided by HEAL-Link Greece.

Declarations

Conflicts of interest The authors declare no conflict of interest.

References

Abdar M, Pourpanah F, Hussain S, Rezazadegan D, Liu L, Ghavamzadeh M, Fieguth P, Cao X, Khosravi A, Acharya UR, Makarenkov V, Nahavandi S (2021) A review of uncertainty quantification in deep learning: techniques, applications and challenges. Information Fusion 76:243–297. https://doi.org/10.1016/j.inffus.2021.05.008

Adam C, Gijbels I (2022) Local polynomial expectile regression. Ann Inst Stat Math 74(2):341–378. https://doi.org/10.1007/s10463-021-00799-y

Ahmad T, Madonski R, Zhang D, Huang C, Mujeeb A (2022) Data-driven probabilistic machine learning in sustainable smart energy/smart energy systems: key developments, challenges, and future research opportunities in the context of smart grid paradigm. Renew Sustain Energy Rev 160:112128. https://doi.org/10.1016/j.rser.2022.112128

Ahmed M, Maume-Deschamps V, Ribereau P (2021) Recognizing a spatial extreme dependence structure: a deep learning approach. Environmetrics. https://doi.org/10.1002/env.2714

Albarello D, D'Amico V (2015) Scoring and testing procedures devoted to probabilistic seismic hazard assessment. Surv Geophys 36(2):269–293. https://doi.org/10.1007/s10712-015-9316-4

Alessandrini S, Delle Monache L, Sperati S, Cervone G (2015) An analog ensemble for short-term probabilistic solar power forecast. Appl Energy 157:95–110. https://doi.org/10.1016/j.apenergy.2015.08.011

Alexandrov A, Benidis K, Bohlke-Schneider M, Flunkert V, Gasthaus J, Januschowski T, Maddix DC, Rangapuram S, Salinas D, Schulz J, Stella L, Türkmen AC, Wang Y (2020) Gluonts: probabilistic and neural time series modeling in Python. J Mach Learn Res 21(116):1–6

Alkhayat G, Mehmood R (2021) A review and taxonomy of wind and solar energy forecasting methods based on deep learning. Energy and AI 4:100060. https://doi.org/10.1016/j.egyai.2021.100060

Antorán J, Allingham JU, Hernández-Lobato JM (2020) Depth uncertainty in neural networks. Adv Neural Inf Process Syst 33:10620–10634

Askanazi R, Diebold FX, Schorfheide F, Shin M (2018) On the comparison of interval forecasts. J Time Ser Anal 39(6):953–965. https://doi.org/10.1111/jtsa.12426

Athey S, Tibshirani J, Wager S (2019) Generalized random forests. Ann Stat 47(2):1179–1203. https://doi.org/10.1214/18-AOS1709

Banerjee S, Gelfand AE, Finley AO, Sang H (2008) Gaussian predictive process models for large spatial data sets. J R Stat Soc: Ser B 70(4):825–848. https://doi.org/10.1111/j.1467-9868.2008.00663.x

Baran S (2014) Probabilistic wind speed forecasting using Bayesian model averaging with truncated normal components. Comput Stat Data Anal 75:227–238. https://doi.org/10.1016/j.csda.2014.02.013

Barbieri MM (2015) Posterior predictive distribution. In: Balakrishnan N, Colton T, Everitt B, Piegorsch W, Ruggeri F, Teugels JL (eds) Wiley StatsRef: Statistics https://doi.org/10.1002/9781118445112.stat07839.

Barczy M (2022) A new example for a proper scoring rule. Commun Stat—Theory Methods 51(11):3705–3712. https://doi.org/10.1080/03610926.2020.1801737

Bassetti F, Casarin R, Ravazzolo F (2018) Bayesian nonparametric calibration and combination of predictive distributions. J Am Stat Assoc 113(522):675–685. https://doi.org/10.1080/01621459.2016.1273117

Bassetti F, Casarin R, Ravazzolo F (2020) Density forecasting. In: Fuleky P (ed) Macroeconomic forecasting in the era of big data. Springer, Cham, pp 465–494

Baumann PFM, Hothorn T, Rügamer D (2021) Deep conditional transformation models. Mach Learn Knowl Discov Databases. https://doi.org/10.1007/978-3-030-86523-8_1

Bellini F, Klar B, Müller A, Rosazza Gianin E (2014) Generalized quantiles as risk measures. Insurance: Math Economics 54(1):41–48. https://doi.org/10.1016/j.insmatheco.2013.10.015

Belloni A, Chernozhukov V, Kato K (2017) High dimensional quantile regression. In: Koenker R, Chernozhukov V, He X, Peng L (eds) Handbook of quantile regression. Chapman and Hall/CRC, New York, pp 253–272

Bentzien S, Friederichs P (2014) Decomposition and graphical portrayal of the quantile score. Q J R Meteorol Soc 140(683):1924–1934. https://doi.org/10.1002/qj.2284

Bermúdez JD, Corberán-Vallet A, Vercher E (2009) Multivariate exponential smoothing: a Bayesian forecast approach based on simulation. Math Comput Simul 79(5):1761–1769. https://doi.org/10.1016/j.matcom.2008.09.004

Bermúdez JD, Segura JV, Vercher E (2010) Bayesian forecasting with the Holt-Winters model. J Operat Res Soc 61(1):164–171. https://doi.org/10.1057/jors.2008.152

Bernardo JM, Smith AFM (2008). Bayesian Theory. https://doi.org/10.1002/9780470316870

Berrisch J, Ziel F (2021) CRPS learning. J Econometrics. https://doi.org/10.1016/j.jeconom.2021.11.008

Bessac J, Naveau P (2021) Forecast score distributions with imperfect observations. Adv Stat Climatol Meteorol Oceanogr 7(2):53–71. https://doi.org/10.5194/ascmo-7-53-2021
Bhat HS, Kumar N, Vaz GJ (2015) Towards scalable quantile regression trees. IEEE Int Conf Big Data 2015:53–60. https://doi.org/10.1109/BigData.2015.7363741
Bickel PJ, Li B (2006) Regularization in Statistics TEST 15:271–344. https://doi.org/10.1007/BF02607055
Billheimer D (2019) Predictive inference and scientific reproducibility. Am Stat 73(sup1):291–295. https://doi.org/10.1080/00031305.2018.1518270
Binois M, Gramacy RB (2021) hetGP: Heteroskedastic Gaussian process modeling and sequential design in R. J Stat Softw 98(13):1–44. https://doi.org/10.18637/jss.v098.i13
Binois M, Gramacy RB, Ludkovski M (2018) Practical heteroscedastic Gaussian process modeling for large simulation experiments. J Comput Graph Stat 27(4):808–821. https://doi.org/10.1080/10618600.2018.1458625
Bjerregård MB, Møller JK, Madsen H (2021) An introduction to multivariate probabilistic forecast evaluation. Energy AI 4:100058. https://doi.org/10.1016/j.egyai.2021.100058
Bostrom H, Asker L, Gurung R, Karlsson I, Lindgren T, Papapetrou P (2017) Conformal prediction using random survival forests. 2017 16Th EEEE Int Conf Mach Learn App. https://doi.org/10.1109/ICMLA.2017.00-57
Bouallègue ZB, Haiden T, Richardson DS (2018) The diagonal score: definition, properties, and interpretations. Quart J R Stat Soc 144(714):1463–1473. https://doi.org/10.1002/qj.3293
Breckling J, Chambers R (1988) M-quantiles. Biometrika 75(4):761–771. https://doi.org/10.1093/biomet/75.4.761
Brehmer JR, Gneiting T (2020) Properization: constructing proper scoring rules via Bayes acts. Ann Inst Stat Math 72(3):659–673. https://doi.org/10.1007/s10463-019-00705-7
Brehmer JR, Gneiting T (2021) Scoring interval forecasts: equal-tailed, shortest, and modal interval. Bernoulli 27(3):1993–2010. https://doi.org/10.3150/20-BEJ1298
Brehmer JR, Strokorb K (2019) Why scoring functions cannot assess tail properties. Electron J Stat 13(2):4015–4034. https://doi.org/10.1214/19-EJS1622
Breiman L (2001a) Random forests. Mach Learn 45(1):5–32. https://doi.org/10.1023/A:1010933404324
Breiman L (2001b) Statistical modeling: the two cultures. Stat Sci 16(3):199–231. https://doi.org/10.1214/ss/1009213726
Brier GW (1950) Verification of forecasts expressed in terms of probability. Mon Weather Rev 78(1):1–3. https://doi.org/10.1175/1520-0493(1950)078%3c0001:VOFEIT%3e2.0.CO;2
Briseño Sanchez G, Hohberg M, Groll A, Kneib T (2020) Flexible instrumental variable distributional regression. J R Stat Soc A Stat Soc 183(4):1553–1574. https://doi.org/10.1111/rssa.12598
Bröcker J (2009) Reliability, sufficiency, and the decomposition of proper scores. Q J R Meteorol Soc 135(643):1512–1519. https://doi.org/10.1002/qj.456
Bröcker J (2012) Evaluating raw ensembles with the continuous ranked probability score. Q J R Meteorol Soc 138(667):1611–1617. https://doi.org/10.1002/qj.1891
Brockhaus S, Rügamer D, Greven S (2020) Boosting functional regression models with FDboost. J Stat Softw 94(10):1–50. https://doi.org/10.18637/jss.v094.i10
Cannon AJ (2011) Quantile regression neural networks: implementation in R and application to precipitation downscaling. Comput Geosci 37(9):1277–1284. https://doi.org/10.1016/j.cageo.2010.07.005
Cannon AJ (2012) Neural networks for probabilistic environmental prediction: Conditional Density Estimation Network Creation and Evaluation (CaDENCE) in R. Comput Geosci 41:126–135. https://doi.org/10.1016/j.cageo.2011.08.023
Carvalho A (2016) An overview of applications of proper scoring rules. Decis Anal 13(4):223–242. https://doi.org/10.1287/deca.2016.0337
Casarin R, Mantoan G, Ravazzolo F (2016) Bayesian calibration of generalized pools of predictive distributions. Econometrics 4(1):17. https://doi.org/10.3390/econometrics4010017
Casati B, Wilson LJ, Stephenson DB, Nurmi P, Ghelli A, Pocernich M, Damrath U, Ebert EE, Brown G, Mason S (2008) Forecast verification: current status and future directions. Meteorol Appl 15(1):3–18. https://doi.org/10.1002/met.52
Chan W-S (1999) Exact joint forecast regions for vector autoregressive models. J Appl Stat 26(1):35–44. https://doi.org/10.1080/02664769922638
Chang B, Joe H (2019) Prediction based on conditional distributions of vine copulas. Comput Stat Data Anal 139:45–63. https://doi.org/10.1016/j.csda.2019.04.015
Chatfield C (1993) Calculating interval forecasts. J Bus Econ Stat 11(2):121–135. https://doi.org/10.1080/07350015.1993.10509938
Chatfield C (1996) Model uncertainty and forecast accuracy. J Forecast 15(7):495–508. https://doi.org/10.1002/(sici)1099-131x(199612)15:7%3c495::aid-for640%3e3.0.co;2-o

Chaudhuri P, Loh W-Y (2002) Nonparametric estimation of conditional quantiles using quantile regression trees. Bernoulli 8(5):561–576
Chen X, Tokdar ST (2021) Joint quantile regression for spatial data. J R Stat Soc: Ser B 83(4):826–852. https://doi.org/10.1111/rssb.12467
Chen Y, Kang Y, Chen Y, Wang Z (2020) Probabilistic forecasting with temporal convolutional neural network. Neurocomputing 399:491–501. https://doi.org/10.1016/j.neucom.2020.03.011
Chen T, Guestrin C (2016) XGBoost: a scalable tree boosting system. In: KDD '16: proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pp. 785–794. https://doi.org/10.1145/2939672.2939785.
Chernozhukov V (2005) Extremal quantile regression. Ann Stat 33(2):806–839. https://doi.org/10.1214/009053604000001165
Chernozhukov V, Fernández-Val I, Galichon A (2010) Quantile and probability curves without crossing. Econometrica 78(3):1093–1125. https://doi.org/10.3982/ECTA7880
Chew V (1966) Confidence, prediction, and tolerance regions for the multivariate normal distribution. J Am Stat Assoc 61(315):605–617. https://doi.org/10.1080/01621459.1966.10480892
Chipman HA, George EI, McCulloch RE (2010) BART: Bayesian additive regression trees. Ann Appl Stat 6(1):266–298. https://doi.org/10.1214/09-AOAS285
Christensen HM (2015) Decomposition of a new proper score for verification of ensemble forecasts. Mon Weather Rev 143(5):1517–1532. https://doi.org/10.1175/MWR-D-14-00150.1
Christensen HM, Moroz IM, Palmer TN (2015) Evaluation of ensemble forecast uncertainty using a new proper score: application to medium-range and seasonal forecasts. Q J R Meteorol Soc 141(687):538–549. https://doi.org/10.1002/qj.2375
Christoffersen PF (1998) Evaluating interval forecasts. Int Econ Rev 39(4):841–862. https://doi.org/10.2307/2527341
Chudý M, Karmakar S, Wu WB (2020) Long-term prediction intervals of economic time series. Empirical Econ 58(1):191–222. https://doi.org/10.1007/s00181-019-01689-2
Čížek P, Sadıkoğlu S (2020) Robust nonparametric regression: a review. Wiley Interdisc Rev. https://doi.org/10.1002/wics.1492
Clarke B, Clarke J (2012) Prediction in several conventional contexts. Stat Surv 6:1–73. https://doi.org/10.1214/12-SS100
Clements MP, Harvey DI (2011) Combining probability forecasts. Int J Forecast 27(2):208–223. https://doi.org/10.1016/j.ijforecast.2009.12.016
Clements MP, Kim JH (2007) Bootstrap prediction intervals for autoregressive time series. Comput Stat Data Anal 51(7):3580–3594. https://doi.org/10.1016/j.csda.2006.09.012
Corani G, Benavoli A, Zaffalon M (2021) Time series forecasting with Gaussian processes needs priors. Mach Learn Knowl Discov Databases. https://doi.org/10.1007/978-3-030-86514-6_7
Cuaresma JC, Feldkircher M, Huber F (2016) Forecasting with global vector autoregressive models: a Bayesian approach. J Appl Economet 31(7):1371–1391. https://doi.org/10.1002/jae.2504
Daouia A, Girard S, Stupfler G (2018) Estimation of tail risk based on extreme expectiles. J R Stat Soc: Ser B 80(2):263–292. https://doi.org/10.1111/rssb.12254
Davis RA, Nielsen MS (2020) Modeling of time series using random forests: theoretical developments. Electronic Journal of Statistics 14(2):3644–3671. https://doi.org/10.1214/20-EJS1758
Dawid AP (1984) Statistical theory: the prequential approach. J R Stat Soc A Stat Soc 147:278–290. https://doi.org/10.2307/2981683
Dawid AP (2007) The geometry of proper scoring rules. Ann Inst Stat Math 59(1):77–93. https://doi.org/10.1007/s10463-006-0099-8
Dawid AP, Musio M (2014) Theory and applications of proper scoring rules. METRON 72(2):169–183. https://doi.org/10.1007/s40300-014-0039-y
De Luna X (2000) Prediction intervals based on autoregression forecasts. J R Stat Soc: Ser D 49(1):87–93. https://doi.org/10.1111/1467-9884.00222
De Backer M, El Ghouch A, Van Keilegom I (2017) Semiparametric copula quantile regression for complete or censored data. Electron J Stat 11(1):1660–1698. https://doi.org/10.1214/17-EJS1273
De Bastiani F, Rigby RA, Stasinopoulous DM, Cysneiros AHMA, Uribe-Opazo MA (2018) Gaussian Markov random field spatial models in GAMLSS. J Appl Stat 45(1):168–186. https://doi.org/10.1080/02664763.2016.1269728
Dearborn K, Frongillo R (2020) On the indirect elicitability of the mode and modal interval. Ann Inst Stat Math 72(5):1095–1108. https://doi.org/10.1007/s10463-019-00719-1
Demut R, Holeňa M (2012) Conformal sets in neural network regression. In: Proceedings of the conference on theory and practice of information technologies (ITAT 2012), pp. 17–24.

Dette H, Van Hecke R, Volgushev S (2014) Some comments on copula-based regression. J Am Stat Assoc 109(507):1319–1324. https://doi.org/10.1080/01621459.2014.916577
Diks C, Panchenko V, Van Dijk D (2011) Likelihood-based scoring rules for comparing density forecasts in tails. J Econometrics 163(2):215–230. https://doi.org/10.1016/j.jeconom.2011.04.001
Diquigiovanni J, Fontana M, Vantini S (2022) Conformal prediction bands for multivariate functional data. J Multivar Anal 189:104879. https://doi.org/10.1016/j.jmva.2021.104879
Du H (2021) Beyond strictly proper scoring rules: the importance of being local. Weather Forecast 36(2):457–468. https://doi.org/10.1175/WAF-D-19-0205.1
Duan LL, Szczesniak RD, Wang X (2017) Functional inverted Wishart for Bayesian multivariate spatial modeling with application to regional climatology model data. Environmetrics. https://doi.org/10.1002/env.2467
Duan T, Avati A, Ding DY, Thai KK, Basu S, Ng A, Schuler A (2020) NGBoost: natural gradient boosting for probabilistic prediction. Proc Mach Learn Res 119:2690–2700
Dunsmore IR (1968) A Bayesian approach to calibration. J R Stat Soc: Ser B 30(2):396–405. https://doi.org/10.1111/j.2517-6161.1968.tb00740.x
Duran RE (2008) Probabilistic sales forecasting for small and medium-size business operations. In: Prasad B (ed) Soft computing applications in business: studies in fuzziness and soft computing, vol 230. Springer, Berlin, Heidelberg
Durham G, Geweke J, Porter-Hudak S, Sowell F (2019) Bayesian inference for ARFIMA models. J Time Ser Anal 40(4):388–410. https://doi.org/10.1111/jtsa.12443
Eaton ML, Giovagnoli A, Sebastiani P (1996) A predictive approach to the Bayesian design problem with application to normal regression models. Biometrika 83(1):111–125. https://doi.org/10.1093/biomet/83.1.111
Efron B (1979) Bootstrap methods: another look at the jackknife. Ann Stat 7(1):1–26. https://doi.org/10.1214/aos/1176344552
Ehm W, Gneiting T (2012) Local proper scoring rules of order two. Ann Stat 40(1):609–637. https://doi.org/10.1214/12-AOS973
Ehm W, Gneiting T, Jordan A, Krüger F (2016) Of quantiles and expectiles: consistent scoring functions, Choquet representations and forecast rankings. J R Stat Soc: Ser B 78(3):505–562. https://doi.org/10.1111/rssb.12154
Eidsvik J, Finley AO, Banerjee S, Rue H (2012) Approximate Bayesian inference for large spatial datasets using predictive process models. Comput Stat Data Anal 56(6):1362–1380. https://doi.org/10.1016/j.csda.2011.10.022
Emmer S, Kratz M, Tasche D (2015) What is the best risk measure in practice? A comparison of standard measures. J Risk 18(2):31–60. https://doi.org/10.21314/JOR.2015.318
Engle RF, Manganelli S (2004) CAViaR: Conditional autoregressive value at risk by regression quantiles. J Bus Econ Stat 22(4):367–381. https://doi.org/10.1198/073500104000000370
Epstein E (1969) A scoring system for probability forecasts of ranked categories. J Appl Meteorol Climatol 8(6):985–987. https://doi.org/10.1175/1520-0450(1969)008%3c0985:ASSFPF%3e2.0.CO;2
Fahrmeir L, Kneib T (2009) Propriety of posteriors in structured additive regression models: Theory and empirical evidence. J Stat Planning and Inference 139(3):843–859. https://doi.org/10.1016/j.jspi.2008.05.036
Fahrmeir L, Kneib T, Lang S, Marx B (2013) Regression: models, methods and applications. Springer, Berlin, Heidelberg
Farooq M, Steinwart I (2017) An SVM-like approach for expectile regression. Comput Stat Data Anal 109:159–181. https://doi.org/10.1016/j.csda.2016.11.010
Fasiolo M, Wood SN, Zaffran M, Nedellec R, Goude Y (2021) qgam: Bayesian nonparametric quantile regression modeling in R. J Stat Softw 100(9):1–31. https://doi.org/10.18637/JSS.V100.I09
Fernández-Delgado M, Sirsat MS, Cernadas E, Alawadi S, Barro S, Febrero-Bande M (2019) An extensive experimental survey of regression methods. Neural Netw 111:11–34. https://doi.org/10.1016/j.neunet.2018.12.010
Ferro CAT (2017) Measuring forecast performance in the presence of observation error. Q J R Meteorol Soc 143(708):2665–2676. https://doi.org/10.1002/qj.3115
Finley AO, Banerjee S, Gelfand AE (2015) spBayes for large univariate and multivariate point-referenced spatio-temporal data models. J Stat Softw 63(13):1–28. https://doi.org/10.18637/jss.v063.i13
Firpo S, Galvao AF, Pinto C, Poirier A, Sanroman G (2021) GMM quantile regression. J Econometrics. https://doi.org/10.1016/j.jeconom.2020.11.014
Fissler T, Frongillo R, Hlavinová J, Rudloff B (2021) Forecast evaluation of quantiles, prediction intervals, and other set-valued functionals. Electronic J Stat 15(1):1034–1084. https://doi.org/10.1214/21-EJS1808

Fong E, Holmes CC (2021) Conformal Bayesian computation. Adv Neural Inf Process Syst 34:18268–18279
Fouedjio F, Klump J (2019) Exploring prediction uncertainty of spatial data in geostatistical and machine learning approaches. Environ Earth Sci. https://doi.org/10.1007/s12665-018-8032-z
Fragoso TM, Bertoli W, Louzada F (2018) Bayesian model averaging: a systematic review and conceptual classification. Int Stat Rev 86(1):1–28. https://doi.org/10.1111/insr.12243
Frazier DT, Maneesoonthorn W, Martin GM, McCabe BPM (2019) Approximate Bayesian forecasting. Int J Forecast 35(2):521–539. https://doi.org/10.1016/j.ijforecast.2018.08.003
Fresoli D (2022) Bootstrap VAR forecasts: the effect of model uncertainties. J Forecast 41(2):279–293. https://doi.org/10.1002/for.2809
Fresoli D, Ruiz E, Pascual L (2015) Bootstrap multi-step forecasts of non-Gaussian VAR models. Int J Forecast 31(3):834–848. https://doi.org/10.1016/j.ijforecast.2014.04.001
Friedberg R, Tibshirani J, Athey S, Wager S (2020) Local linear forests. J Comput Graph Stat 30(2):503–517. https://doi.org/10.1080/10618600.2020.1831930
Friederichs P, Thorarinsdottir TL (2012) Forecast verification for extreme value distributions with an application to probabilistic peak wind prediction. Environmetrics 23(7):579–594. https://doi.org/10.1002/env.2176
Friedman JH (2001) Greedy function approximation: a gradient boosting machine. Ann Stat 29(5):1189–1232. https://doi.org/10.1214/aos/1013203451
Friedman JH (2020) Contrast trees and distribution boosting. Proc Natl Acad Sci USA 117(35):21175–21184. https://doi.org/10.1073/pnas.1921562117
Gaba A, Tsetlin I, Winkler RL (2017) Combining interval forecasts. Decis Anal 14(1):1–20. https://doi.org/10.1287/deca.2016.0340
Gal Y, Ghahramani Z (2016) Dropout as a Bayesian approximation: representing model uncertainty in deep learning. Proc Mach Learn Res 48:1050–1059
Gammerman A (2012) Conformal predictors: progress in Artificial. Intelligence 1:203–204. https://doi.org/10.1007/s13748-012-0024-8
Gandy A, Jana K, Veraart AED (2022) Scoring predictions at extreme quantiles. AStA Adv Stat Anal. https://doi.org/10.1007/s10182-021-00421-9
Gasthaus J, Benidis K, Wang Y, Rangapuram SS, Salinas D, Flunkert V, Januschowski T (2020) Probabilistic forecasting with spline quantile function RNNs. Proc Mach Learn Res 89:1901–1910
Gawlikowski J, Tassi CRN, Ali M, Lee J, Humt M, Feng J, Kruspe A, Triebel R, Jung P, Roscher R, Shahzad M, Yang W, Bamler R, Zhu XX (2023) A survey of uncertainty in deep neural networks. Artif Intell Rev. https://doi.org/10.1007/s10462-023-10562-9
Geisser S (1965) Bayesian estimation in multivariate analysis. Ann Math Stat 36(1):150–159. https://doi.org/10.1214/aoms/1177700279
Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB (2013) Bayesian data analysis, 3rd edn. Chapman and Hall/CRC
Genest C, Zidek JV (1986) Combining probability distributions: a critique and an annotated bibliography. Stat Sci 1(1):114–148. https://doi.org/10.1214/ss/1177013825
Gensler A, Sick B, Vogt S (2018) A review of uncertainty representations and metaverification of uncertainty assessment techniques for renewable energies. Renew Sustain Energy Rev 96:352–379. https://doi.org/10.1016/j.rser.2018.07.042
Geraci M, Bottai M (2014) Linear quantile mixed models. Stat Comput 24(3):461–479. https://doi.org/10.1007/s11222-013-9381-9
Geweke J, Whiteman C (2006) Chapter 1 Bayesian Forecasting. In: Elliott G, Granger CWJ, Timmermann A (eds) Handbook of economic forecasting, vol 1. Elsevier, pp 3–80
Ghil M, Yiou P, Hallegatte S, Malamud BD, Naveau P, Soloviev A, Friederichs P, Keilis-Borok V, Kondrashov D, Kossobokov V, Mestre O, Nicolis C, Rust HW, Shebalin P, Vrac M, Witt A, Zaliapin I (2011) Extreme events: dynamics, statistics and prediction. Nonlinear Process Geophys 18(3):295–350. https://doi.org/10.5194/npg-18-295-2011
Ghosh S, Vera Liao Q, Ramamurthy KN, Navratil J, Sattigeri P, Varshney K, Zhang Y (2022) Uncertainty quantification 360: a hands-on tutorial. In: CODS-COMAD 2022: 5th joint international conference on data science & management of data (9th ACM IKDD CODS and 27th COMAD), pp. 333–335. https://doi.org/10.1145/3493700.3493767.
Girard A, Rasmussen CE, Candela JQ, Murray-Smith R (2003) Gaussian process priors with uncertain inputs application to multiple-step ahead time series forecasting. Adv Neural Inf Process Syst 15:545–552
Girard S, Stupfler G, Usseglio-Carleve A (2022a) Functional estimation of extreme conditional expectiles. Econometrics and Statistics 21:131–158. https://doi.org/10.1016/j.ecosta.2021.05.006

Girard S, Stupfler G, Usseglio-Carleve A (2022b) Nonparametric extreme conditional expectile estimation. Scand J Stat 49(1):78–115. https://doi.org/10.1111/sjos.12502
Gneiting T (2011a) Making and evaluating point forecasts. J Am Stat Assoc 106(494):746–762. https://doi.org/10.1198/jasa.2011.r10138
Gneiting T (2011b) Quantiles as optimal point forecasts. Int J Forecast 27(2):197–207. https://doi.org/10.1016/j.ijforecast.2009.12.015
Gneiting T, Katzfuss M (2014) Probabilistic forecasting. Ann Rev Stat Its App 1:125–151. https://doi.org/10.1146/annurev-statistics-062713-085831
Gneiting T, Raftery AE (2007) Strictly proper scoring rules, prediction, and estimation. J Am Stat Assoc 102(477):359–378. https://doi.org/10.1198/016214506000001437
Gneiting T, Ranjan R (2011) Comparing density forecasts using threshold-and quantile-weighted scoring rules. J Bus Econ Stat 29(3):411–422. https://doi.org/10.1198/jbes.2010.08110
Gneiting T, Ranjan R (2013) Combining predictive distributions. Electron J Stat 7(1):1747–1782. https://doi.org/10.1214/13-EJS823
Gneiting T, Raftery AE, Westveld AH III, Goldman T (2005) Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. Mon Weather Rev 133(5):1098–1118. https://doi.org/10.1175/MWR2904.1
Gneiting T, Balabdaoui F, Raftery AE (2007) Probabilistic forecasts, calibration and sharpness. J R Stat Soc: Ser B 69(2):243–268. https://doi.org/10.1111/j.1467-9868.2007.00587.x
Good IJ (1952) Rational decisions. J R Stat Soc: Ser B 14(1):107–114. https://doi.org/10.1111/j.2517-6161.1952.tb00104.x
Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative adversarial nets. Adv Neural Inf Process Syst 27:2672–2680
Gourieroux C, Jasiak J (2008) Dynamic quantile models. J Econometrics 147(1):198–205. https://doi.org/10.1016/j.jeconom.2008.09.028
Grant A, Johnstone D, Kwon OK (2019) A probability scoring rule for simultaneous events. Decis Anal 16(4):301–313. https://doi.org/10.1287/deca.2019.0393
Green PJ (2013) Discussion of 'beyond mean regression.' Stat Model 13(4):305–315. https://doi.org/10.1177/1471082X13494160
Greven S, Scheipl F (2017) A general framework for functional regression modelling. Stat Model 17(1–2):1–35. https://doi.org/10.1177/1471082X16681317
Grigoletto M (2005) Bootstrap prediction regions for multivariate autoregressive processes. Stat Methods Appl 14(2):179–207. https://doi.org/10.1007/s10260-005-0113-y
Grimit EP, Gneiting T, Berrocal VJ, Johnson NA (2006) The continuous ranked probability score for circular variables and its application to mesoscale forecast ensemble verification. Q J R Meteorol Soc 132(621C):2925–2942. https://doi.org/10.1256/qj.05.235
Größer J, Okhrin O (2021) Copulae: an overview and recent developments. Wiley Interdisc Rev 14(3):e1557. https://doi.org/10.1002/wics.1557
Grushka-Cockayne Y, Jose VRR (2020) Combining prediction intervals in the M4 competition. Int J Forecast 36(1):178–185. https://doi.org/10.1016/j.ijforecast.2019.04.015
Guerbyenne H, Hamdi F (2015) Bootstrapping periodic state-space models. Commun Stat—Simulation Comput 44(2):374–401. https://doi.org/10.1080/03610918.2013.777737
Hall SG, Mitchell J (2007) Combining density forecasts. Int J Forecast 23(1):1–13. https://doi.org/10.1016/j.ijforecast.2006.08.001
Hallin M, Šiman M (2017) Multiple output quantile regression. In: Koenker R, Chernozhukov V, He X, Peng L (eds) Handbook of quantile regression. Chapman and Hall/CRC, New York, pp 185–207
Hallin M, Lu Z, Yu K (2009) Local linear spatial quantile regression. Bernoulli 15(3):659–686. https://doi.org/10.3150/08-BEJ168
Hallin M, Paindaveine D, Šiman M (2010) Multivariate quantiles and multiple-output regression quantiles: from L1 optimization to halfspace depth. Ann Stat 38(2):635–6 69. https://doi.org/10.1214/09-AOS723
Hamori S, Motegi K, Zhang Z (2020) Copula-based regression models with data missing at random. J Multivar Anal 180:104654. https://doi.org/10.1016/j.jmva.2020.104654
Hansen BE (2006) Interval forecasts and parameter uncertainty. J Econometrics 135(1–2):377–398. https://doi.org/10.1016/j.jeconom.2005.07.030
Hariri RH, Fredericks EM, Bowers KM (2019) Uncertainty in big data analytics: survey, opportunities, and challenges. J Big Data. https://doi.org/10.1186/s40537-019-0206-3
Harva M (2007) A variational EM approach to predictive uncertainty. Neural Netw 20(4):550–558. https://doi.org/10.1016/j.neunet.2007.04.010

Hastie T, Tibshirani R (1986) Generalized additive models (with discussion). Stat Sci 1(3):297–310. https://doi.org/10.1214/ss/1177013604
Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning. Springer, New York
Haupt SE, Gagne DJ, Hsieh WW, Krasnopolsky V, McGovern A, Marzban C, Moninger W, Lakshmanan V, Tissot P, Williams JK (2022) The history and practice of AI in the environmental sciences. Bull Am Meteor Soc 103(5):E1351–E1370. https://doi.org/10.1175/BAMS-D-20-0234.1
He X, Ng P (1999) Quantile splines with several covariates. J Stat Plan Inference 75(2):343–352. https://doi.org/10.1016/S0378-3758(98)00153-0
He Y, Zhang X, Zhang L (2018) Variable selection for high dimensional Gaussian copula regression model: an adaptive hypothesis testing procedure. Comput Stat Data Anal 124:132–150. https://doi.org/10.1016/j.csda.2018.03.003
He XD, Kou S, Peng X (2022) Risk measures: robustness, elicitability, and backtesting. Annu Rev Stat Its App 9:141–166. https://doi.org/10.1146/annurev-statistics-030718-105122
Hefley TJ, Hooten MB, Hanks EM, Russell RE, Walsh DP (2017) Dynamic spatio-temporal models for spatial data. Spatial Statistics 20:206–220. https://doi.org/10.1016/j.spasta.2017.02.005
Hengl T, Nussbaum M, Wright MN, Heuvelink GBM, Gräler B (2018) Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. PeerJ. https://doi.org/10.7717/peerj.5518
Henzi A, Kleger G-R, Ziegel JF (2021a) Distributional (single) index models. J Am Stat Assoc. https://doi.org/10.1080/01621459.2021.1938582
Henzi A, Ziegel JF, Gneiting T (2021b) Isotonic distributional regression. J R Stat Soc: Ser B 83(5):963–993. https://doi.org/10.1111/rssb.12450
Hernández B, Raftery AE, Pennington SR, Parnell AC (2018) Bayesian additive regression trees using Bayesian model averaging. Stat Comput 28(4):869–890. https://doi.org/10.1007/s11222-017-9767-1
Hersbach H (2000) Decomposition of the continuous ranked probability score for ensemble prediction systems. Weather Forecast 15(5):559–570. https://doi.org/10.1175/1520-0434(2000)015%3c0559:DOTCRP%3e2.0.CO;2
Hewamalage H, Bergmeir C, Bandara K (2021) Recurrent neural networks for time series forecasting: current status and future directions. Int J Forecast 37(1):388–427. https://doi.org/10.1016/j.ijforecast.2020.06.008
Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735
Hoerl AE, Kennard RW (1970) Ridge regression: Biased estimation for nonorthogonal problems. Technometrics 12(1):55–67. https://doi.org/10.1080/00401706.1970.10488634
Hoeting JA, Madigan D, Raftery AE, Volinsky CT (1999) Bayesian model averaging: a tutorial. Stat Sci 14(4):382–401. https://doi.org/10.1214/ss/1009212519
Hofner B, Mayr A, Robinzonov N, Schmid M (2014) Model-based boosting in R: a hands-on tutorial using the R package mboost. Comput Stat 29(1–2):3–35. https://doi.org/10.1007/s00180-012-0382-5
Hofner B, Mayr A, Schmid M (2016) GamboostLSS: an R package for model building and variable selection in the GAMLSS framework. J Stat Softw. https://doi.org/10.18637/jss.v074.i01
Hong T, Fan S (2016) Probabilistic electric load forecasting: a tutorial review. Int J Forecast 32(3):914–938. https://doi.org/10.1016/j.ijforecast.2015.11.011
Hora SC, Kardeş E (2015) Calibration, sharpness and the weighting of experts in a linear opinion pool. Ann Operat Res 229(1):429–450. https://doi.org/10.1007/s10479-015-1846-0
Hothorn T (2020a) Most likely transformations: the mlt package. J Stat Softw 92(1):1–68. https://doi.org/10.18637/jss.v092.i01
Hothorn T (2020b) Transformation boosting machines. Stat Comput 30(1):141–152. https://doi.org/10.1007/s11222-019-09870-4
Hothorn T, Zeileis A (2021) Predictive distribution modeling using transformation forests. J Comput Graph Stat 30(4):1181–1196. https://doi.org/10.1080/10618600.2021.1872581
Hothorn T, Kneib T, Bühlmann P (2014) Conditional transformation models. J R Stat Soc: Ser B 76(1):3–27. https://doi.org/10.1111/rssb.12017
Hothorn T, Möst L, Bühlmann P (2018) Most likely transformations. Scand J Stat 45(1):110–134. https://doi.org/10.1111/sjos.12291
Hu T, Guo Q, Li Z, Shen X, Sun H (2020) Distribution-free probability density forecast through deep neural networks. IEEE Trans Neural Netw Learn Syst 31(2):612–625. https://doi.org/10.1109/TNNLS.2019.2907305
Huang Z, Zhao T (2022) Predictive performance of ensemble hydroclimatic forecasts: verification metrics, diagnostic plots and forecast attributes. Wiley Interdiscip Rev Water 9(2):e1580. https://doi.org/10.1002/wat2.1580

Hüllermeier E, Waegeman W (2021) Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. Mach Learn 110(3):457–506. https://doi.org/10.1007/s10994-021-05946-3
Huser R (2021) Editorial: EVA 2019 data competition on spatio-temporal prediction of Red Sea surface temperature extremes. Extremes 24(1):91–104. https://doi.org/10.1007/s10687-019-00369-9
Hyndman RJ, Koehler AB (2006) Another look at measures of forecast accuracy. Int J Forecast 22(4):679–688. https://doi.org/10.1016/j.ijforecast.2006.03.001
Hyndman RJ, Shang HL (2009) Forecasting functional time series. J Korean Stat Soc 38(3):199–211. https://doi.org/10.1016/j.jkss.2009.06.002
Hyndman RJ, Bashtannyk DM, Grunwald GK (1996) Estimating and visualizing conditional densities. J Comput Graph Stat 5(4):315–336. https://doi.org/10.1080/10618600.1996.10474715
Hyvärinen A, Dayan P (2005) Estimation of non-normalized statistical models by score matching. J Mach Learn Res 6(24):695–709
Iacopini M, Ravazzolo F, Rossini L (2022) Proper scoring rules for evaluating density forecasts with asymmetric loss functions. J Bus Econ Stat. https://doi.org/10.1080/07350015.2022.2035229
Ibacache-Pulgar G, Paula GA, Cysneiros FJA (2013) Semiparametric additive models under symmetric distributions. TEST 22(1):103–121. https://doi.org/10.1007/s11749-012-0309-z
Jankowiak M, Pleiss G, Gardner JR (2020) Parametric Gaussian process regressors. Proc Mach Learn Res 119:4702–4712
Jantre SR, Bhattacharya S, Maiti T (2021) Quantile regression neural networks: a Bayesian approach. J Stat Plan Inference. https://doi.org/10.1007/s42519-021-00189-w
Januschowski T, Wang Y, Torkkola K, Erkkilä T, Hasson H, Gasthaus J (2021) Forecasting with trees. Int J Forecast. https://doi.org/10.1016/j.ijforecast.2021.10.004
Jia Y, Jeong J-H (2022) Deep learning for quantile regression under right censoring: DeepQuantreg. Comput Stat Data Anal 165:107323. https://doi.org/10.1016/j.csda.2021.107323
Jiang C, Jiang M, Xu Q, Huang X (2017) Expectile regression neural network model with applications. Neurocomputing 247:73–86. https://doi.org/10.1016/j.neucom.2017.03.040
Jiang Y, Lin F, Zhou Y (2021) The kth power expectile regression. Ann Inst Stat Math 73(1):83–113. https://doi.org/10.1007/s10463-019-00738-y
Johansson U, Boström H, Löfström T, Linusson H (2014) Regression conformal prediction with random forests. Mach Learn 97(1–2):155–176. https://doi.org/10.1007/s10994-014-5453-0
Johansson U, Linusson H, Löfström T, Boström H (2018) Interpretable regression trees using conformal prediction. Expert Syst Appl 97:394–404. https://doi.org/10.1016/j.eswa.2017.12.041
Johnstone DJ, Jose VRR, Winkler RL (2011) Tailored scoring rules for probabilities. Decis Anal 8(4):256–268. https://doi.org/10.1287/deca.1110.0216
Jones MC (1994) Expectiles and *M*-quantiles are quantiles. Statist Probab Lett 20(2):149–153. https://doi.org/10.1016/0167-7152(94)90031-0
Jordan A, Krüger F, Lerch S (2019) Evaluating probabilistic forecasts with scoringRules. J Stat Softw 90(12):1–37. https://doi.org/10.18637/jss.v090.i12
Juutilainen I, Tamminen S, Röning J (2012) Exceedance probability score: a novel measure for comparing probabilistic predictions. J Stat Plan Inference 6(3):452–467. https://doi.org/10.1080/15598608.2012.695663
Kabir HMD, Khosravi A, Hosen MA, Nahavandi S (2018) Neural network-based uncertainty quantification: a survey of methodologies and applications. IEEE Access 6:36218–36234. https://doi.org/10.1109/ACCESS.2018.2836917
Kapetanios G, Mitchell J, Price S, Fawcett N (2015) Generalised density forecast combinations. J Econometrics 188(1):150–165. https://doi.org/10.1016/j.jeconom.2015.02.047
Kaplan D (2021) On the quantification of model uncertainty: a Bayesian perspective. Psychometrika 86(1):215–238. https://doi.org/10.1007/s11336-021-09754-5
Kaplan D, Yavuz S (2020) An approach to addressing multiple imputation model uncertainty using Bayesian model averaging. Multivar Behav Res 55(4):553–567. https://doi.org/10.1080/00273171.2019.1657790
Katzfuss M, Cressie N (2012) Bayesian hierarchical spatio-temporal smoothing for very large datasets. Environmetrics 23(1):94–107. https://doi.org/10.1002/env.1147
Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu T-Y (2017) LightGBM: a highly efficient gradient boosting decision tree. Adv Neural Inf Process Syst 30:3146–3154
Kelly R, Chen K (2022) Distribution free prediction intervals for multiple functional regression. Stat Its Interface 15(2):161–170. https://doi.org/10.4310/20-SII646
Khajeh H, Laaksonen H (2022) Applications of probabilistic forecasting in smart grids: a review. Appl Sci 12(4):1823. https://doi.org/10.3390/app12041823

Khosravi A, Nahavandi S, Creighton D, Atiya AF (2011) Comprehensive review of neural network-based prediction intervals and new advances. IEEE Trans Neural Networks 22(9):1341–1356. https://doi.org/10.1109/TNN.2011.2162110

Khosravi A, Nahavandi S, Creighton D (2013) A neural network-GARCH-based method for construction of prediction intervals. Electr Power Syst Res 96:185–193. https://doi.org/10.1016/j.epsr.2012.11.007

Kitahara D, Leng K, Tezuka Y, Hirabayashi A (2021) Simultaneous spline quantile regression under shape constraints. In: 2020 28th European Signal Processing Conference (EUSIPCO), pp. 2423–2427. https://doi.org/10.23919/Eusipco47968.2020.9287462

Kleiber W, Raftery AE, Baars J, Gneiting T, Mass CF, Grimit E (2011) Locally calibrated probabilistic temperature forecasting using geostatistical model averaging and local Bayesian model averaging. Mon Weather Rev 139(8):2630–2649. https://doi.org/10.1175/2010MWR3511.1

Klein N, Kneib T (2016) Simultaneous inference in structured additive conditional copula regression models: a unifying Bayesian approach. Stat Comput 26(4):841–860. https://doi.org/10.1007/s11222-015-9573-6

Klein N, Kneib T, Klasen S, Lang S (2015a) Bayesian structured additive distributional regression for multivariate responses. J R Stat Soc: Ser C 64(4):569–591. https://doi.org/10.1111/rssc.12090

Klein N, Kneib T, Lang S (2015b) Bayesian generalized additive models for location, scale, and shape for zero-inflated and overdispersed count data. J Am Stat Assoc 110(509):405–419. https://doi.org/10.1080/01621459.2014.912955

Klein N, Nott DJ, Smith MS (2021) Marginally calibrated deep distributional regression. J Comput Graph Stat 30(2):467–483. https://doi.org/10.1080/10618600.2020.1807996

Kneib T (2013) Beyond mean regression. Stat Model 13(4):275–303. https://doi.org/10.1177/1471082X13494159

Kneib T, Silbersdorff A, Säfken B (2021) Rage against the mean—a review of distributional regression approaches. Econometrics Stat. https://doi.org/10.1016/j.ecosta.2021.07.006

Knüppel M, Krüger F (2022) Forecast uncertainty, disagreement, and the linear pool. J Appl Economet 37(1):23–41. https://doi.org/10.1002/jae.2834

Koenker R (2017) Quantile regression: 40 years on. Annu Rev Econmics 9:155–176. https://doi.org/10.1146/annurev-economics-063016-103651

Koenker RW, Bassett G Jr (1978) Regression quantiles. Econometrica 46(1):33–50. https://doi.org/10.2307/1913643

Koenker R, Hallock KF (2001) Quantile regression. J Econ Perspect 15(4):143–156. https://doi.org/10.1257/jep.15.4.143

Koenker R, Xiao Z (2006) Quantile autoregression. J Am Stat Assoc 101(475):980–990. https://doi.org/10.1198/016214506000000672

Koenker R, Chernozhukov V, He X, Peng L (2017) Handbook of quantile regression. Chapman and Hall/CRC, New York

Kolev N, Paiva D (2009) Copula-based regression models: a survey. J Stat Plan Inference 139(11):3847–3856. https://doi.org/10.1016/j.jspi.2009.05.023

Koltchinskii VI (1997) M-estimation, convexity and quantiles. Ann Stat 25(2):435–477. https://doi.org/10.1214/aos/1031833659

Kompa B, Snoek J, Beam AL (2021) Empirical frequentist coverage of deep learning uncertainty quantification procedures. Entropy 23(12):1608. https://doi.org/10.3390/e23121608

Koochali A, Schichtel P, Dengel A, Ahmed S (2019) Probabilistic forecasting of sensory data with generative adversarial networks—ForGAN. IEEE Access 7:63868–63880. https://doi.org/10.1109/ACCESS.2019.2915544

Koochali A, Dengel A, Ahmed S (2021) If you like it, GAN it—probabilistic multivariate times series forecast with GAN [†]. Eng Proc 5(1):40. https://doi.org/10.3390/engproc2021005040

Kook L, Baumann PFM, Dürr O, Sick B, Rügamer D (2023) Estimating conditional distributions with neural networks using R package deeptrafo. https://arxiv.org/abs/2211.13665

Kraus D, Czado C (2017) D-vine copula based quantile regression. Comput Stat Data Anal 110:1–18. https://doi.org/10.1016/j.csda.2016.12.009

Krüger F, Lerch S, Thorarinsdottir T, Gneiting T (2021) Predictive inference based on Markov Chain Monte Carlo output. Int Stat Rev 89:274–301. https://doi.org/10.1111/insr.12405

Kuan C-M, Yeh J-H, Hsu YC (2009) Assessing value at risk with CARE, the conditional autoregressive expectile models. J Econometrics 150(2):261–270. https://doi.org/10.1016/j.jeconom.2008.12.002

Kuleshov V, Fenner N, Ermon S (2018) Accurate uncertainties for deep learning using calibrated regression. Proc Mach Learn Res 80:2796–2804

Kupiec PH (1995) Techniques for verifying the accuracy of risk measurement models. J Derivatives 3(2):73–84. https://doi.org/10.3905/jod.1995.407942
Lai TL, Gross ST, Shen DB (2011) Evaluating probability forecasts. Ann Stat 39(5):2356–2382. https://doi.org/10.1214/11-AOS902
Lakshminarayanan B, Pritzel A, Blundell C (2017) Simple and scalable predictive uncertainty estimation using deep ensembles. Adv Neural Inf Process Syst 30:6402–6413
Lampinen J, Vehtari A (2001) Bayesian approach for neural networks—review and case studies. Neural Netw 14(3):257–274. https://doi.org/10.1016/S0893-6080(00)00098-8
Landon J, Singpurwalla ND (2008) Choosing a coverage probability for prediction intervals. Am Stat 62(2):120–124. https://doi.org/10.1198/000313008X304062
Lang MN, Schlosser L, Hothorn T, Mayr GJ, Stauffer R, Zeileis A (2020) Circular regression trees and forests with an application to probabilistic wind direction forecasting. J R Stat Soc: Ser C 69(5):1357–1374. https://doi.org/10.1111/rssc.12437
Lecun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521(7553):436–444. https://doi.org/10.1038/nature14539
Lee HKH (2000) Consistency of posterior distributions for neural networks. Neural Netw 13(6):629–642. https://doi.org/10.1016/S0893-6080(00)00045-9
Lee YS, Scholtes S (2014) Empirical prediction intervals revisited. Int J Forecast 30(2):217–234. https://doi.org/10.1016/j.ijforecast.2013.07.018
Lei J, Wasserman L (2014) Distribution-free prediction bands for non-parametric regression. J R Stat Soc: Ser B 76(1):71–96. https://doi.org/10.1111/rssb.12021
Lerch S, Thorarinsdottir TL, Ravazzolo F, Gneiting T (2017) Forecaster's dilemma: extreme events and forecast evaluation. Stat Sci 32(1):106–127. https://doi.org/10.1214/16-STS588
Leung TY, Leutbecher M, Reich S, Shepherd TG (2021) Forecast verification: relating deterministic and probabilistic metrics. Q J R Meteorol Soc 147(739):3124–3134. https://doi.org/10.1002/qj.4120
Li J (2011) Bootstrap prediction intervals for SETAR models. Int J Forecast 27(2):320–332. https://doi.org/10.1016/j.ijforecast.2010.01.013
Li R, Peng L (2017) Assessing quantile prediction with censored quantile regression models. Biometrics 73(2):517–528. https://doi.org/10.1111/biom.12627
Li G, Li Y, Tsai C-L (2015) Quantile correlations and quantile autoregressive modeling. J Am Stat Assoc 110(509):246–261. https://doi.org/10.1080/01621459.2014.892007
Li W, Duan Q, Miao C, Ye A, Gong W, Di Z (2017) A review on statistical postprocessing methods for hydrometeorological ensemble forecasting. Wiley Interdisc Rev: Water. https://doi.org/10.1002/wat2.1246
Li R, Reich BJ, Bondell HD (2021) Deep distribution regression. Comput Stat Data Anal 159:107203. https://doi.org/10.1016/j.csda.2021.107203
Liang WWJ, Lee HKH (2019) Bayesian nonstationary Gaussian process models via treed process convolutions. Adv Data Anal Classif 13(3):797–818. https://doi.org/10.1007/s11634-018-0341-2
Liao L, Park C, Choi H (2019) Penalized expectile regression: an alternative to penalized quantile regression. Ann Inst Stat Math 71(2):409–438. https://doi.org/10.1007/s10463-018-0645-1
Lichtendahl KC Jr, Winkler RL (2007) Probability elicitation, scoring rules, and competition among forecasters. Manage Sci 53(11):1745–1755. https://doi.org/10.1287/mnsc.1070.0729
Lichtendahl KC Jr, Grushka-Cockayne Y, Winkler RL (2013) Is it better to average probabilities or quantiles? Manage Sci 59(7):1594–1611. https://doi.org/10.1287/mnsc.1120.1667
Liu S-I (1994) Multiperiod Bayesian forecasts for AR models. Ann Inst Stat Math 46(3):429–452. https://doi.org/10.1007/BF00773509
Liu S-I (1995) Bayesian multiperiod forecasts for ARX models. Ann Inst Stat Math 47(2):211–224. https://doi.org/10.1007/BF00773458
Liu S, Li S (2022) Multi-model D-vine copula regression model with vine copula-based dependence description. Comput Chem Eng 161:107788. https://doi.org/10.1016/j.compchemeng.2022.107788
Liu JZ, Lin Z, Padhy S, Tran D, Bedrax-Weiss T, Lakshminarayanan B (2020) Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. Adv Neural Inf Process Syst 33 (NeurIPS 2020).
Loaiza-Maya R, Martin GM, Frazier DT (2021) Focused Bayesian prediction. J Appl Economet 36(5):517–543. https://doi.org/10.1002/jae.2810
Lorenz EN (1969) Atmospheric predictability as revealed by naturally occurring analogues. J Atmos Sci 26(4):636–646. https://doi.org/10.1175/1520-0469(1969)26%3c636:APARBN%3e2.0.CO;2
Lu B, Hardin J (2021) A unified framework for random forest prediction error estimation. J Mach Learn Res 22(8):1–41
Lu W, Zhu Z, Lian H (2020) High-dimensional quantile tensor regression. J Mach Learn Res 21(250):1–31

Machete RL (2013) Contrasting probabilistic scoring rules. J Stat Plan Inference 143(10):1781–1790. https://doi.org/10.1016/j.jspi.2013.05.012
MacNab YC (2018) Some recent work on multivariate Gaussian Markov random fields. TEST 27(3):497–541. https://doi.org/10.1007/s11749-018-0605-3
Makridakis S, Spiliotis E, Assimakopoulos V, Chen Z, Gaba A, Tsetlin I, Winkler RL (2021) The M5 uncertainty competition: results, findings and conclusions. Int J Forecast. https://doi.org/10.1016/j.ijforecast.2021.10.009
Mancini T, Calvo-Pardo H, Olmo J (2021) Extremely randomized neural networks for constructing prediction intervals. Neural Netw 144:113–128. https://doi.org/10.1016/j.neunet.2021.08.020
Marra G, Radice R (2017) Bivariate copula additive models for location, scale and shape. Comput Stat Data Anal 112:99–113. https://doi.org/10.1016/j.csda.2017.03.004
Masarotto G, Varin C (2017) Gaussian copula regression in R. J Stat Softw. Do: https://doi.org/10.18637/jss.v077.i08.
Matheson JE, Winkler RL (1976) Scoring rules for continuous probability distributions. Manage Sci 22(10):1087–1096. https://doi.org/10.1287/mnsc.22.10.1087
Mathonsi T, Van Zyl TL (2020) Prediction interval construction for multivariate point forecasts using deep learning. In: 2020 7th International Conference on Soft Computing & Machine Intelligence (ISCMI), pp 88–95. https://doi.org/10.1109/ISCMI51676.2020.9311603.
Mayr A, Hofner B (2018) Boosting for statistical modelling—a non-technical introduction. Stat Model 18(3–4):365–384. https://doi.org/10.1177/1471082X17748086
Mayr A, Fenske N, Hofner B, Kneib T, Schmid M (2012) Generalized additive models for location, scale and shape for high dimensional data-a flexible approach based on boosting. J R Stat Soc: Ser C 61(3):403–427. https://doi.org/10.1111/j.1467-9876.2011.01033.x
Mayr A, Binder H, Gefeller O, Schmid M (2014a) The evolution of boosting algorithms: from machine learning to statistical modelling. Methods Inf Med 53(06):419–427. https://doi.org/10.3414/ME13-01-0122
Mayr A, Binder H, Gefeller O, Schmid M (2014b) Extending statistical boosting: an overview of recent methodological developments. Methods Inf Med 53(06):428–435. https://doi.org/10.3414/ME13-01-0123
Mayr A, Hofner B, Waldmann E, Hepp T, Meyer S, Gefeller O (2017) An update on statistical boosting in biomedicine. Comp Math Methods Med 2017:6083072. https://doi.org/10.1155/2017/6083072
Meinshausen N (2006) Quantile regression forests. J Mach Learn Res 7(35):983–999
Merkle EC, Steyvers M (2013) Choosing a strictly proper scoring rule. Decis Anal 10(4):292–304. https://doi.org/10.1287/deca.2013.0280
Messner JW, Mayr GJ, Zeileis A (2016) Heteroscedastic censored and truncated regression with crch. R J 8(1):173–181. https://doi.org/10.32614/rj-2016-012
Michaelis P, Klein N, Kneib T (2018) Bayesian multivariate distributional regression with skewed responses and skewed random effects. J Comput Graph Stat 27(3):602–611. https://doi.org/10.1080/10618600.2017.1395343
Mizera I (2017) Quantile regression: penalized. In: Koenker R, Chernozhukov V, He X, Peng L (eds) Handbook of quantile regression. Chapman and Hall/CRC, New York, pp 21–40
Monahan JF (1983) Fully Bayesian analysis of ARMA time series models. J Econometrics 21(3):307–331. https://doi.org/10.1016/0304-4076(83)90048-9
Moon SJ, Jeon J-J, Lee JSH, Kim Y (2021) Learning multiple quantiles with neural networks. J Comput Graph Stat 30(4):1238–1248. https://doi.org/10.1080/10618600.2021.1909601
Mukhopadhyay S, Wang K (2020) Breiman's "Two Cultures" revisited and reconciled. https://arxiv.org/abs/2005.13596
Müller P, West M, Maceachern S (1997) Bayesian models for non-linear autoregressions. J Time Ser Anal 18(6):593–614. https://doi.org/10.1111/1467-9892.00070
Murphy AH, Daan H (1985) Forecast evaluation. In: Murphy AH, Katz RW (eds) Probability, statistics and decision making in the atmospheric sciences. Westview Press, Boulder, pp 379–437
Nelsen RB (2006) An introduction to copulas. Springer, New York
Newey WK, Powell JL (1987) Asymmetric least squares estimation and testing. Econometrica 55(4):819–847. https://doi.org/10.2307/1911031
Nguyen T-T, Huang JZ, Nguyen TT (2015) Two-level quantile regression forests for bias correction in range prediction. Mach Learn 101(1–3):325–343. https://doi.org/10.1007/s10994-014-5452-1
Nix DA, Weigend AS (1994) Estimating the mean and variance of the target probability distribution. Proc 1994 EEE Int Conf Neural Netw 1:55–60. https://doi.org/10.1109/ICNN.1994.374138
Noh H, El Ghouch A, Bouezmarni T (2013) Copula-based regression estimation and inference. J Am Stat Assoc 108(502):676–688. https://doi.org/10.1080/01621459.2013.783842

Noh H, Ghouch AE, Van Keilegom I (2015) Semiparametric conditional quantile estimation through copula-based multivariate models. J Bus Econ Stat 33(2):167–178. https://doi.org/10.1080/07350015.2014.926171
Nott D (2006) Semiparametric estimation of mean and variance functions for non-Gaussian data. Comput Stat 21(3–4):603–620. https://doi.org/10.1007/s00180-006-0017-9
Nowotarski J, Weron R (2018) Recent advances in electricity price forecasting: a review of probabilistic forecasting. Renew Sustain Energy Rev 81(1):1548–1568. https://doi.org/10.1016/j.rser.2017.05.234
Oesting M, Schlather M, Friederichs P (2017) Statistical post-processing of forecasts for extremes using bivariate Brown–Resnick processes with an application to wind gusts. Extremes 20(2):309–332. https://doi.org/10.1007/s10687-016-0277-x
Opschoor A, van Dijk D, van der Wel M (2017) Combining density forecasts using focused scoring rules. J Appl Economet 32(7):1298–1313. https://doi.org/10.1002/jae.2575
Ord K, Lowe S (1996) Automatic forecasting. Am Stat 50(1):88–94. https://doi.org/10.1080/00031305.1996.10473549
Ovadia Y, Fertig E, Ren J, Nado Z, Sculley D, Nowozin S, Dillon JV, Lakshminarayanan B, Snoek J (2019) Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. Adv Neural Inf Process Syst 32:13991–14002
Pacchiardi L, Adewoyin R, Dueben P, Dutta R (2022) Probabilistic forecasting with generative networks via scoring rule minimization. https://arxiv.org/abs/2112.08217
Padoan SA, Stupfler G (2022) Joint inference on extreme expectiles for multivariate heavy-tailed distributions. Bernoulli 28(2):1021–1048. https://doi.org/10.3150/21-BEJ1375
Pai JS, Ravishanker N (1996) Bayesian modelling of ARFIMA processes by Markov Chain Monte Carlo methods. J Forecast 15(2):63–82. https://doi.org/10.1002/(SICI)1099-131X(199603)15:2%3c63::AID-FOR606%3e3.0.CO;2-5
Papacharalampous G, Tyralis H (2022) A review of machine learning concepts and methods for addressing challenges in probabilistic hydrological post-processing and forecasting. Front Water 4:961954. https://doi.org/10.3389/frwa.2022.961954
Papacharalampous G, Tyralis H, Langousis A, Jayawardena AW, Sivakumar B, Mamassis N, Montanari A, Koutsoyiannis D (2019) Probabilistic hydrological post-processing at scale: Why and how to apply machine-learning quantile regression algorithms. Water 11(10):2126. https://doi.org/10.3390/w11102126
Papadopoulos H, Vovk V, Gammerman A (2011) Regression conformal prediction with nearest neighbours. J Artif Intell Res 40:815–840. https://doi.org/10.1613/jair.3198
Parry M, Dawid AP, Lauritzen S (2012) Proper local scoring rules. Ann Stat 40(1):561–592. https://doi.org/10.1214/12-AOS971
Patton AJ (2020) Comparing possibly misspecified forecasts. J Bus Econ Stat 38(4):796–809. https://doi.org/10.1080/07350015.2019.1585256
Peng L (2017) Quantile regression for survival analysis. In: Koenker R, Chernozhukov V, He X, Peng L (eds) Handbook of quantile regression. Chapman and Hall/CRC, New York, pp 89–103
Peng L (2021) Quantile regression for survival data. Annu Rev Stat Its App 8:413–437. https://doi.org/10.1146/annurev-statistics-042720-020233
Piironen J, Vehtari A (2017) Comparison of Bayesian predictive methods for model selection. Stat Comput 27(3):711–735. https://doi.org/10.1007/s11222-016-9649-y
Powell JL (1986) Censored regression quantiles. J Economet 32(1):143–155. https://doi.org/10.1016/0304-4076(86)90016-3
Pratola MT, Chipman HA, George EI, McCulloch RE (2020) Heteroscedastic BART via multiplicative regression trees. J Comput Graph Stat 29(2):405–417. https://doi.org/10.1080/10618600.2019.1677243
Prokudin S, Gehler P, Nowozin S (2018) Deep directional statistics: pose estimation with uncertainty quantification. Comput vis: ECCV 2018:542–559. https://doi.org/10.1007/978-3-030-01240-3_33
Raftery AE, Gneiting T, Balabdaoui F, Polakowski M (2005) Using Bayesian model averaging to calibrate forecast ensembles. Mon Weather Rev 133(5):1155–1174. https://doi.org/10.1175/MWR2906.1
Raiffa H, Schlaifer R (1961) Applied Statistical Decision Theory. Colonial Press, Clinton
Ranjan R, Gneiting T (2010) Combining probability forecasts. J R Stat Soc: Ser B 72(1):71–91. https://doi.org/10.1111/j.1467-9868.2009.00726.x
Rasmussen CE (2004) Gaussian Processes in machine learning. In: Bousquet O, von Luxburg U, Rätsch G (eds) Advanced lectures on machine learning. Springer, Berlin, Heidelberg, pp 63–71
Rasmussen CE, Williams CKI (2006) Gaussian processes for machine learning. MIT Press, Cambridge

Rasp S, Lerch S (2018) Neural networks for postprocessing ensemble weather forecasts. Mon Weather Rev 146(11):3885–3900. https://doi.org/10.1175/MWR-D-18-0187.1
Ravishanker N, Ray BK (2002) Bayesian prediction for vector ARFIMA processes. Int J Forecast 18(2):207–214. https://doi.org/10.1016/S0169-2070(01)00153-4
Ray EL, Sakrejda K, Lauer SA, Johansson MA, Reich NG (2017) Infectious disease prediction with kernel conditional density estimation. Stat Med 36(30):4908–4929. https://doi.org/10.1002/sim.7488
Regnier E (2018) Probability forecasts made at multiple lead times. Manage Sci 64(5):2407–2426. https://doi.org/10.1287/mnsc.2016.2720
Rémillard B, Nasri B, Bouezmarni T (2017) On copula-based conditional quantile estimators. Statist Probab Lett 128:14–20. https://doi.org/10.1016/j.spl.2017.04.014
Rigby RA, Stasinopoulos DM (2005) Generalized additive models for location, scale and shape. J R Stat Soc: Ser C 54(3):507–554. https://doi.org/10.1111/j.1467-9876.2005.00510.x
Rigby RA, Stasinopoulos DM (2006) Using the Box-Cox t distribution in GAMLSS to model skewness and kurtosis. Stat Model 6(3):209–229. https://doi.org/10.1191/1471082X06st122oa
Rigby RA, Stasinopoulos DM, Voudouris V (2013) Discussion: a comparison of GAMLSS with quantile regression. Stat Model 13(4):335–348. https://doi.org/10.1177/1471082X13494316
Risser MD, Turek D (2020) Bayesian inference for high-dimensional nonstationary Gaussian processes. J Stat Comput Simulation. https://doi.org/10.1080/00949655.2020.1792472
Robert CP (2007) The Bayesian choice. Springer, New York
Roberts HV (1965) Probabilistic prediction. J Am Stat Assoc 60(309):50–62. https://doi.org/10.1080/01621459.1965.10480774
Roberts S, Osborne M, Ebden M, Reece S, Gibson N, Aigrain S (2013) Gaussian processes for time-series modelling. Philos Trans R Soc A. https://doi.org/10.1098/rsta.2011.0550
Rohekar RY, Gurwicz Y, Nisimov S, Novik G (2019) Modeling uncertainty by learning a hierarchy of deep neural connections. Adv Neural Inf Process Syst 32:4244–4254
Romano Y, Patterson E, Candès EJ (2019) Conformalized quantile regression. Adv Neural Inf Process Syst 32:3543–3553
Rothfuss J, Ferreira F, Walther S, Ulrich M (2019) Conditional density estimation with neural networks: Best practices and benchmarks. https://arxiv.org/abs/1903.00954
Roulston M, Smith L (2003) Combining dynamical and statistical ensembles. Tellus A 55(1):16–30. https://doi.org/10.3402/tellusa.v55i1.12082
Roy M-H, Larocque D (2020) Prediction intervals with random forests. Stat Methods Med Res 29(1):205–229. https://doi.org/10.1177/0962280219829885
Rue H, Held L (2005) Gaussian Markov random fields: theory and applications. Chapman and Hall/CRC, New York
Rue H, Riebler A, Sørbye SH, Illian JB, Simpson DP, Lindgren FK (2017) Bayesian computing with INLA: a review. Annu Rev Stat Its App 4:395–421. https://doi.org/10.1146/annurev-statistics-060116-054045
Rügamer D, Baumann PFM, Kneib T, Hothorn T (2023a) Probabilistic time series forecasts with autoregressive transformation models. Stat Comput. https://doi.org/10.1007/s11222-023-10212-8
Rügamer D, Kolb C, Fritz C, Pfisterer F, Kopper P, Bischl B, Shen R, Bukas C et al (2023b) deepregression: a flexible neural network framework for semi-structured deep distributional regression. J Stat Softw 105(2):1–31. https://doi.org/10.18637/jss.v105.i02
Rügamer D, Kolb C, Klein N (2023c) Semi-structured distributional regression. Am Stat. https://doi.org/10.1080/00031305.2022.2164054
Salinas D, Flunkert V, Gasthaus J, Januschowski T (2020) DeepAR: probabilistic forecasting with autoregressive recurrent networks. Int J Forecast 36(3):1181–1191. https://doi.org/10.1016/j.ijforecast.2019.07.001
Scheipl F, Gertheiss J, Greven S (2016) Generalized functional additive mixed models. Electr J Stat 10(1):1455–1492. https://doi.org/10.1214/16-EJS1145
Scheuerer M, Hamill TM (2015) Variogram-based proper scoring rules for probabilistic forecasts of multivariate quantities. Mon Weather Rev 143(4):1321–1334. https://doi.org/10.1175/MWR-D-14-00269.1
Schlosser L, Hothorn T, Stauffer R, Zeileis A (2019) Distributional regression forests for probabilistic precipitation forecasting in complex terrain. Ann Appl Stat 13(3):1564–1589. https://doi.org/10.1214/19-AOAS1247
Schmid M, Wickler F, Maloney KO, Mitchell R, Fenske N, Mayr A (2013) Boosted beta regression. PLoS ONE 8(4):e61623. https://doi.org/10.1371/journal.pone.0061623
Schmidhuber J (2015) Deep learning in neural networks: an overview. Neural Netw 61:85–117. https://doi.org/10.1016/j.neunet.2014.09.003

Seipp A, Uslar V, Weyhe D, Timmer A, Otto-Sobotka F (2021) Weighted expectile regression for right-censored data. Stat Med 40(25):5501–5520. https://doi.org/10.1002/sim.9137
Sen R, Yu H-F, Dhillon I (2019) Think globally, act locally: A deep neural network approach to high-dimensional time series forecasting. Adv Neural Inf Process Syst 32:4837–4846
Serpell C, Araya I, Valle C, Allende H (2019) Probabilistic forecasting using Monte Carlo dropout neural networks. In: Progress in pattern recognition, image analysis, computer vision, and applications, pp. 387–397. https://doi.org/10.1007/978-3-030-33904-3_36.
Sesia M, Candès EJ (2020) A comparison of some conformal quantile regression methods. Stat 9(1):e261. https://doi.org/10.1002/sta4.261
Sezer OB, Gudelek MU, Ozbayoglu AM (2020) Financial time series forecasting with deep learning: a systematic literature review: 2005–2019. Appl Soft Comput 90:106181. https://doi.org/10.1016/j.asoc.2020.106181
Shafer G, Vovk V (2008) A tutorial on conformal prediction. J Mach Learn Res 9:371–421
Shan K, Yang Y (2009) Combining regression quantile estimators. Stat Sin 19(3):1171–1191
Shang HL, Hyndman RJ (2011) Nonparametric time series forecasting with dynamic updating. Math Comput Simul 81(7):1310–1324. https://doi.org/10.1016/j.matcom.2010.04.027
Shim J, Kim Y, Lee J, Hwang C (2012) Estimating value at risk with semiparametric support vector quantile regression. Comput Statistics 27(4):685–700. https://doi.org/10.1007/s00180-011-0283-z
Shmueli G (2010) To explain or to predict? Stat Sci 25(3):289–310. https://doi.org/10.1214/10-STS330
Sidén P, Lindsten F (2020) Deep Gaussian Markov random fields. Proc Mach Learn Res 119:8916–8926
Sillmann J, Thorarinsdottir T, Keenlyside N, Schaller N, Alexander LV, Hegerl G, Seneviratne SI, Vautard R, Zhang X, Zwiers FW (2017) Understanding, modeling and predicting weather and climate extremes: challenges and opportunities. Weather Clim Extremes 18:65–74. https://doi.org/10.1016/j.wace.2017.10.003
Silva PCL, Sadaei HJ, Guimaraes FG (2016) Interval forecasting with fuzzy time series. IEEE Symp Ser Comput Intell (SSCI) 2016:1–8. https://doi.org/10.1109/SSCI.2016.7850010
Silva PCL, Sadaei HJ, Ballini R, Guimaraes FG (2020) Probabilistic forecasting with fuzzy time series. IEEE Trans Fuzzy Syst 28(8):1771–1784. https://doi.org/10.1109/TFUZZ.2019.2922152
Sims CA, Zha T (1998) Bayesian methods for dynamic multivariate models. Int Econ Rev 39(4):949–968. https://doi.org/10.2307/2527347
Sklar A (1959) Fonctions de répartition à *n* dimensions et leurs marges. Publications Del'institut De Statistique De L'université De Paris 8:229–231
Smith MS, Klein N (2021) Bayesian inference for regression copulas. J Bus Econ Stat 39(3):712–728. https://doi.org/10.1080/07350015.2020.1721295
Smyl S (2020) A hybrid method of exponential smoothing and recurrent neural networks for time series forecasting. Int J Forecast 36(1):75–85. https://doi.org/10.1016/j.ijforecast.2019.03.017
Snyder RD, Ord JK, Koehler AB (2001) Prediction intervals for ARIMA models. J Bus Econ Stat 19(2):217–225. https://doi.org/10.1198/073500101316970430
Sobotka F, Kneib T (2012) Geoadditive expectile regression. Comput Stat Data Anal 56(4):755–767. https://doi.org/10.1016/j.csda.2010.11.015
Sobotka F, Kauermann G, Waltrup LS, Kneib T (2013) On confidence intervals for semiparametric expectile regression. Stat Comput 23(2):135–148. https://doi.org/10.1007/s11222-011-9297-1
Spiegel E, Sobotka F, Kneib T (2017) Model selection in semiparametric expectile regression. Electron J Stat 11(2):3008–3038. https://doi.org/10.1214/17-EJS1307
Spiegel E, Kneib T, Otto-Sobotka F (2020) Spatio-temporal expectile regression models. Stat Model 20(4):386–409. https://doi.org/10.1177/1471082X19829945
Sprangers O, Schelter S, De Rijke M (2021) Probabilistic gradient boosting machines for large-scale probabilistic regression. In: PKDD '21: Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining, pp 1510–1520. https://doi.org/10.1145/3447548.3467278.
Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R (2014) Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 15:1929–1958
Stankevičiūte K, Alaa AM, van der Schaar M (2021) Conformal time-series forecasting. Adv Neural Inf Process Syst 34:6216–6228
Stasinopoulos DM, Rigby RA (2007) Generalized additive models for location scale and shape (GAMLSS) in R. J Stat Softw 23(7):1–46. https://doi.org/10.18637/jss.v023.i07
Stasinopoulos MD, Rigby RA, De Bastiani F (2018) GAMLSS: a distributional regression approach. Stat Model 18(3–4):248–273. https://doi.org/10.1177/1471082X18759144
Staszewska-Bystrova A (2011) Bootstrap prediction bands for forecast paths from vector autoregressive models. J Forecast 30(8):721–735. https://doi.org/10.1002/for.1205

Steel MFJ (2020) Model averaging and its use in economics. J Econ Literature 58(3):644–719. https://doi.org/10.1257/JEL.20191385
Strömer A, Staerk C, Klein N, Weinhold L, Titze S, Mayr A (2022) Deselection of base-learners for statistical boosting—with an application to distributional regression. Stat Methods Med Res 31(2):207–224. https://doi.org/10.1177/09622802211051088
Sungur EA (2005) Some observations on copula regression functions. Commun Stat—Theory Methods 34(9–10):1967–1978. https://doi.org/10.1080/03610920500201244
Swiatkowski J, Roth K, Veeling B, Tran L, Dillon J, Snoek J, Mandt S, Salimans T, Jenatton R, Nowozin S (2019) The *k*-tied normal distribution: a compact parameterization of Gaussian mean field posteriors in Bayesian neural networks. Proc Mach Learn Res 119:9289–9299
Tagasovska N, Lopez-Paz D (2019) Single-model uncertainties for deep learning. Adv Neural Inf Process Syst 32:6417–6428
Taggart R (2022a) Evaluation of point forecasts for extreme events using consistent scoring functions. Q J R Meteorol Soc 148(742):306–320. https://doi.org/10.1002/qj.4206
Taggart RJ (2022b) Point forecasting and forecast evaluation with generalized Huber loss. Electron J Stat 16(1):201–231. https://doi.org/10.1214/21-EJS1957
Tajmouati S, El-Wahbi B, Dakkon M (2022) Applying regression conformal prediction with nearest neighbors to time series data. Commun Stat: Simulation Comput. https://doi.org/10.1080/03610918.2022.2057538
Takeuchi I, Le QV, Sears TD, Smola AJ (2006) Nonparametric quantile estimation. J Mach Learn Res 7(45):1231–1264
Tay AS, Wallis KF (2000) Density forecasting: a survey. J Forecast 19(4):235–254. https://doi.org/10.1002/1099-131x(200007)19:4%3c235::aid-for772%3e3.3.co;2-c
Taylor JW (2000) A quantile regression neural network approach to estimating the conditional density of multiperiod returns. J Forecast 19(4):299–311. https://doi.org/10.1002/1099-131x(200007)19:4%3c299::aid-for775%3e3.3.co;2-m
Taylor JW (2021) Evaluating quantile-bounded and expectile-bounded interval forecasts. Int J Forecast 37(2):800–811. https://doi.org/10.1016/j.ijforecast.2020.09.007
Taylor JW, Bunn DW (1998) Combining forecast quantiles using quantile regression: investigating the derived weights, estimator bias and imposing constraints. J Appl Stat 25(2):193–206. https://doi.org/10.1080/02664769823188
Taylor JW, Bunn DW (1999) Quantile regression approach to generating prediction intervals. Manage Sci 45(2):225–237. https://doi.org/10.1287/mnsc.45.2.225
Taylor SJ, Letham B (2018) Forecasting at scale. Am Stat 72(1):37–45. https://doi.org/10.1080/00031305.2017.1380080
Tepegjozova M, Zhou J, Claeskens G, Czado C (2022) Nonparametric C- and D-vine-based quantile regression. Dependence Modell 10(1):1–21. https://doi.org/10.1515/demo-2022-0100
Teye M, Azizpour H, Smith K (2018) Bayesian uncertainty estimation for batch normalized deep networks. Proc Mach Learn Res 80:4907–4916
Thiagarajan JJ, Venkatesh B, Sattigeri P, Bremer P-T (2020) Building calibrated deep models via uncertainty matching with auxiliary interval predictors. In: The thirty-fourth AAAI Conference on Artificial Intelligence (AAAI-20), pp. 6005–6012. https://doi.org/10.1609/aaai.v34i04.6062.
Thomas J, Mayr A, Bischl B, Schmid M, Smith A, Hofner B (2018) Gradient boosting for distributional regression: faster tuning and improved variable selection via noncyclical updates. Stat Comput 28(3):673–687. https://doi.org/10.1007/s11222-017-9754-6
Thorgeirsson AT, Gauterin F (2021) Probabilistic predictions with federated learning. Entropy 23(1):41. https://doi.org/10.3390/e23010041
Tibshirani R (1996) Regression shrinkage and selection via the lasso. J R Stat Soc: Ser B 58:267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x
Titterington DM (2004) Bayesian methods for neural networks and related models. Stat Sci 19(1):128–139. https://doi.org/10.1214/088342304000000099
Tony Cai T, Zhang L (2018) High-dimensional Gaussian copula regression: adaptive estimation and statistical inference. Stat Sin 28(2):963–993. https://doi.org/10.5705/ss.202016.0041
Torossian L, Picheny V, Faivre R, Garivier A (2020) A review on quantile regression for stochastic computer experiments. Reliab Eng Syst Saf 201:106858. https://doi.org/10.1016/j.ress.2020.106858
Tran D, Ranganath R, Blei DM (2017) Hierarchical implicit models and likelihood-free variational inference. Adv Neural Inf Process Syst 30:5523–5533
Tsallis C (1988) Possible generalization of Boltzmann-Gibbs statistics. J Stat Phys 52:479–487. https://doi.org/10.1007/BF01016429

Tung NT, Huang JZ, Nguyen TT, Khan I (2014) Bias-corrected quantile regression forests for high-dimensional data. Int Conf Mach Learn Cybern 2014:1–6. https://doi.org/10.1109/ICMLC.2014.7009082
Tyralis H, Papacharalampous G (2021) Boosting algorithms in energy research: a systematic review. Neural Comput Appl 33(21):14101–14117. https://doi.org/10.1007/s00521-021-05995-8
Tyralis H, Papacharalampous G, Burnetas A, Langousis A (2019a) Hydrological post-processing using stacked generalization of quantile regression algorithms: large-scale application over CONUS. J Hydrol 577:123957. https://doi.org/10.1016/j.jhydrol.2019.123957
Tyralis H, Papacharalampous G, Langousis A (2019b) A brief review of random forests for water scientists and practitioners and their recent history in water resources. Water 11(5):910. https://doi.org/10.3390/w11050910
Tyralis H, Papacharalampous G, Dogulu N, Chun KP (2023) Deep Huber quantile regression networks. https://arxiv.org/abs/2306.10306
Umlauf N, Kneib T (2018) A primer on Bayesian distributional regression. Stat Model 18(3–4):219–247. https://doi.org/10.1177/1471082X18759140
Umlauf N, Klein N, Zeileis A (2018) BAMLSS: Bayesian additive models for location, scale, and shape (and beyond). J Comput Graph Stat 27(3):612–627. https://doi.org/10.1080/10618600.2017.1407325
Umlauf N, Klein N, Simon T, Zeileis A (2021) bamlss: a Lego toolbox for flexible Bayesian regression (and beyond). J Stat Softw 100(4):1–53. https://doi.org/10.18637/JSS.V100.I04
van der Meer DW, Widén J, Munkhammar J (2018) Review on probabilistic forecasting of photovoltaic power production and electricity consumption. Renew Sustain Energy Rev 81(1):1484–1512. https://doi.org/10.1016/j.rser.2017.05.212
Vannitsem S, Bremnes JB, Demaeyer J, Evans GR, Flowerdew J, Hemri S, Lerch S, Roberts N, Theis S, Atencia A, Bouallègue ZB, Bhend J, Dabernig M, De Cruz L, Hieta L, Mestre O, Moret L, Plenković IO, Schmeits M, Taillardat M, Van den Bergh J, Van Schaeybroeck B, Whan K, Ylhaisi J (2021) Statistical postprocessing for weather forecasts: review, challenges, and avenues in a big data world. Bull Am Meteor Soc 102(3):E681–E699. https://doi.org/10.1175/BAMS-D-19-0308.1
Vasiloudis T, de Francisci MG, Boström H (2019) Quantifying uncertainty in online regression forests. J Mach Learn Res 20(155):1–35
Vehtari A, Ojanen J (2012) A survey of Bayesian predictive methods for model assessment, selection and comparison. Stat Surv 6(1):142–228. https://doi.org/10.1214/12-ss102
Vehtari A, Gelman A, Gabry J (2017) Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. Stat Comput 27(5):1413–1432. https://doi.org/10.1007/s11222-016-9696-4
Virbickaite A, Ausín MC, Galeano P (2015) Bayesian inference methods for univariate and multivariate GARCH models: a survey. Journal of Economic Surveys 29(1):76–96. https://doi.org/10.1111/joes.12046
Vovk V, Gammerman A, Shafer G (2005) Algorithmic learning in a random world. Springer, New York
Vrontos ID, Dellaportas P, Politis DN (2000) Full Bayesian inference for GARCH and EGARCH models. J Bus Econ Stat 18(2):187–198. https://doi.org/10.1080/07350015.2000.10524861
Waldmann E (2018) Quantile regression: a short story on how and why. Stat Model 18(3–4):203–218. https://doi.org/10.1177/1471082X18759142
Waldmann E, Kneib T, Yue YR, Lang S, Flexeder C (2013) Bayesian semiparametric additive quantile regression. Stat Model 13(3):223–252. https://doi.org/10.1177/1471082X13480650
Waldmann E, Sobotka F, Kneib T (2017) Bayesian regularisation in geoadditive expectile regression. Stat Comput 27(6):1539–1553. https://doi.org/10.1007/s11222-016-9703-9
Waltrup LS, Sobotka F, Kneib T, Kauermann G (2015) Expectile and quantile regression—David and Goliath? Stat Model 15(5):433–456. https://doi.org/10.1177/1471082X14561155
Wang L (2017) Nonconvex penalized quantile regression: a review of methods, theory and algorithms. In: Koenker R, Chernozhukov V, He X, Peng L (eds) Handbook of quantile regression. Chapman and Hall/CRC, New York, pp 273–292
Wang F, Gelfand AE (2014) Modeling space and space-time directional data using projected Gaussian processes. J Am Stat Assoc 109(508):1565–1580. https://doi.org/10.1080/01621459.2014.934454
Wang HJ, Li D (2013) Estimation of extreme conditional quantiles through power transformation. J Am Stat Assoc 108(503):1062–1074. https://doi.org/10.1080/01621459.2013.820134
Wang HJ, Yang Y (2017) Bayesian quantile regression. In: Koenker R, Chernozhukov V, He X, Peng L (eds) Handbook of quantile regression. Chapman and Hall/CRC, New York, pp 41–54
Wang HJ, Li D, He X (2012) Estimation of high conditional quantiles for heavy-tailed distributions. J Am Stat Assoc 107(500):1453–1464. https://doi.org/10.1080/01621459.2012.716382
Wei Y, Carroll RJ (2009) Quantile regression with measurement error. J Am Stat Assoc 104(487):1129–1143. https://doi.org/10.1198/jasa.2009.tm08420

Weinhold L, Schmid M, Mitchell R, Maloney KO, Wright MN, Berger M (2020) A random forest approach for bounded outcome variables. J Comput Graph Stat 29(3):639–658. https://doi.org/10.1080/10618600.2019.1705310
Weron R (2014) Electricity price forecasting: a review of the state-of-the-art with a look into the future. Int J Forecast 30(4):1030–1081. https://doi.org/10.1016/j.ijforecast.2014.08.008
Wilks DS (2011) Chapter 8—Forecast verification. In: Wilks DS (ed.) International Geophysics, vol. 100, pp. 301–394. https://doi.org/10.1016/B978-0-12-385022-5.00008-7.
Winkler RL (1972) A decision-theoretic approach to interval estimation. J Am Stat Assoc 67(337):187–191. https://doi.org/10.1080/01621459.1972.10481224
Winkler RL (1996) Scoring rules and the evaluation of probabilities (with discussion and reply). TEST 5(1):1–60. https://doi.org/10.1007/BF02562681
Winkler RL, Murphy AH (1968) "Good" probability assessors. J Appl Meteorol Climatol 7(5):751–758. https://doi.org/10.1175/1520-0450(1968)007%3c0751:PA%3e2.0.CO;2
Winkler RL, Grushka-Cockayne Y, Lichtendahl KC, Jose VRR (2019) Probability forecasts and their combination: a research perspective. Decis Anal 16(4):239–260. https://doi.org/10.1287/deca.2019.0391
Wolpert DH (1992) Stacked generalization. Neural Netw 5(2):241–259. https://doi.org/10.1016/S0893-6080(05)80023-1
Wood SN (2020) Inference and computation with generalized additive models and their extensions. TEST 29(2):307–339. https://doi.org/10.1007/s11749-020-00711-5
Wu JJ (2012) Semiparametric forecast intervals. J Forecast 31(3):189–228. https://doi.org/10.1002/for.1185
Wu D, Gao L, Chinazzi M, Xiong X, Vespignani A, Ma Y-A, Yu R (2021) Quantifying uncertainty in deep spatiotemporal forecasting. In: PKDD '21: proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining, pp 1841–1851. https://doi.org/10.1145/3447548.3467325
Xiao Z, Koenker R (2009) Conditional quantile estimation for generalized autoregressive conditional heteroscedasticity models. J Am Stat Assoc 104(488):1696–1712. https://doi.org/10.1198/jasa.2009.tm09170
Xie Z, Wen H (2019) Composite quantile regression long short-term memory network. In: Artificial neural networks and machine learning—ICANN 2019: text and time series, pp. 513–524. https://doi.org/10.1007/978-3-030-30490-4_41
Xu SG, Reich BJ (2021) Bayesian nonparametric quantile process regression and estimation of marginal quantile effects. Biometrics. https://doi.org/10.1111/biom.13576
Xu Q, Liu X, Jiang C, Yu K (2016) Quantile autoregression neural network model with applications to evaluating value at risk. Appl Soft Comput 49:1–12. https://doi.org/10.1016/j.asoc.2016.08.003
Xu Q, Deng K, Jiang C, Sun F, Huang X (2017) Composite quantile regression neural network with applications. Expert Syst Appl 76:129–139. https://doi.org/10.1016/j.eswa.2017.01.054
Xu Q, Liu S, Jiang C, Zhuo X (2021) QRNN-MIDAS: a novel quantile regression neural network for mixed sampling frequency data. Neurocomputing 457:84–105. https://doi.org/10.1016/j.neucom.2021.06.006
Yang D, van der Meer D (2021) Post-processing in solar forecasting: ten overarching thinking tools. Renew Sustain Energy Rev 140:110735. https://doi.org/10.1016/j.rser.2021.110735
Yang D, Wang W, Gueymard CA, Hong T, Kleissl J, Huang J, Perez MJ, Perez R, Bright JM, Xia X, van der Meer D, Peters IM (2022) A review of solar forecasting, its dependence on atmospheric sciences and implications for grid integration: towards carbon neutrality. Renew Sustain Energy Rev 161:112348. https://doi.org/10.1016/j.rser.2022.112348
Yao Y, Vehtari A, Simpson D, Gelman A (2018) Using stacking to average Bayesian predictive distributions. Bayesian Anal 13(3):917–1003. https://doi.org/10.1214/17-BA1091
Ye SS, Padilla OHM (2021) Non-parametric quantile regression via the k-nn fused lasso. J Mach Learn Res 22(111):1–38
Ying Z, Sit T (2017) Survival analysis: a quantile perspective. In: Koenker R, Chernozhukov V, He X, Peng L (eds) Handbook of quantile regression. Chapman and Hall/CRC, New York, pp 69–87
Yu K, Moyeed RA (2001) Bayesian quantile regression. Statist Probab Lett 54(4):437–447. https://doi.org/10.1016/S0167-7152(01)00124-9
Yu L, Yang Z, Tang L (2018) Quantile estimators with orthogonal pinball loss function. J Forecast 37(3):401–417. https://doi.org/10.1002/for.2510
Yuan S (2015) Random gradient boosting for predicting conditional quantiles. J Stat Comput Simul 85(18):3716–3726. https://doi.org/10.1080/00949655.2014.1002099
Zammit-Mangion A, Ng TLJ, Vu Q, Filippone M (2021) Deep compositional spatial models. J Am Stat Assoc. https://doi.org/10.1080/01621459.2021.1887741

Zhang Y, Nadarajah S (2018) A review of backtesting for value at risk. Commun Stat—Theory Methods 47(15):3616–3639. https://doi.org/10.1080/03610926.2017.1361984
Zhang Y, Wang J, Wang X (2014) Review on probabilistic forecasting of wind power generation. Renew Sustain Energy Rev 32:255–270. https://doi.org/10.1016/j.rser.2014.01.033
Zhang L, Datta A, Banerjee S (2019) Practical Bayesian modeling and inference for massive spatial data sets on modest computing environments. Stat Anal Data Mining 12(3):197–209. https://doi.org/10.1002/sam.11413
Zhang H, Zimmerman J, Nettleton D, Nordman DJ (2020) Random forest prediction intervals. Am Stat 74(4):392–406. https://doi.org/10.1080/00031305.2019.1585288
Zhao J, Zhang Y (2018) Variable selection in expectile regression. Commun Stat—Theory Methods 47(7):1731–1746. https://doi.org/10.1080/03610926.2017.1324989
Zhao J, Chen Y, Zhang Y (2018) Expectile regression for analyzing heteroscedasticity in high dimension. Statist Probab Lett 137:304–311. https://doi.org/10.1016/j.spl.2018.02.006
Zhao Y, Gijbels I, Van Keilegom I (2020) Inference for semiparametric Gaussian copula model adjusted for linear regression using residual ranks. Bernoulli 26(4):2815–2846. https://doi.org/10.3150/20-BEJ1208
Zheng S (2011) Gradient descent algorithms for quantile regression with smooth approximation. Int J Mach Learn Cybern 2(3):191–207. https://doi.org/10.1007/s13042-011-0031-2
Zhou X, Liu H, Pourpanah F, Zeng T, Wang X (2022) A survey on epistemic (model) uncertainty in supervised learning: recent advances and applications. Neurocomputing. https://doi.org/10.1016/j.neucom.2021.10.119
Ziel F (2021) M5 competition uncertainty: overdispersion, distributional forecasting, GAMLSS, and beyond. Int J Forecast. https://doi.org/10.1016/j.ijforecast.2021.09.008
Ziel F, Steinert R (2018) Probabilistic mid- and long-term electricity price forecasting. Renew Sustain Energy Rev 94:251–266. https://doi.org/10.1016/j.rser.2018.05.038