Scientific computing for probabilistic metrics

STING engineer's internship report

ECN supervisor:
Anthony NOUY

Internship supervisore:
Josselin Le Gal La Salle

STING 2023

JIAN Fan   2nd year of engineering studies

Dates : du 11 avril 2023 au 31 août 2023

Address : 40 avenue de Soweto 97455 Saint-Pierre.

August 24, 2023

Contents

Remerciements

Je voudrais remercier toutes les personnes qui ont contribué à mon stage de quatre mois et demi. Je suis très heureuse d'avoir choisi de passer du temps dans le laboratoire Piment à la fin de mes deux années d'études en France.

Merci à mon maître de stage,M.Josselin Le Gal La Salle, qui m'a expliqué les concepts dès le début et m'a permis d'entrer rapidement dans le sujet.Merci beaucoup à Rodrigo et Mathieu pour leur aide avant et après le stage, Ce stage a confirmé mon intérêt pour le travail dans le domaine de la prévision de l'énergie solaire et de l'intégration de l'énergie. Je suis sûre que j'aurai d'autres occasions d'être exposée à des sujets connexes au cours de mes études et de ma vie de chercheuse.

Je remercie tout particulièrement mon tuteur à l'ECN, M.Anthony Nouy, qui a enseigné les cours de probabilités(MNP) à l'école et m'a donné les connaissances préalables nécessaires pour m'aider dans ce stage. Je suis très heureuse d'avoir choisi l'option Mathématiques appliquées à l'ECN et de l'avoir achevée avec satisfaction, ce qui m'aidera plus tard à travailler dans le secteur de système électrique en Chine.

Merci à tous les stagiaires et à tous les professeurs de l'ECN qui m'ont enseigné, je vous souhaite le meilleur.

Acknowledgements

I would like to thank everyone who contributed to my four-and-a-half-month internship. I am very happy to have chosen to spend time in the Piment laboratory at the end of my two years of study in France.

Thanks to my internship supervisor Josslin, who explained the concepts to me right from the start and got me quickly into the subject. Thanks to Rodrigo and Mathieu for their help before and after the internship, this internship confirmed my interest in working in the field of solar energy forecasting and energy integration. I'm sure I'll have other opportunities to be exposed to related topics during my studies and my life as a researcher.

Special thanks go to my tutor at ECN, Anthony Nouy, who taught the numerical probability method at the school and gave me the necessary background knowledge to help me with this internship. I'm very happy to have chosen the Applied Mathematics option at ECN and to have completed it with satisfaction, which will help me later on to work in the power system sector in China.

Thank you to all the interns and teachers at ECN who taught me, I wish you all the best.

1. Introduction

1.1 Presentation of the laboratory Piment

PIMENT was formed in January 2010 from the integration of several research teams covering a wide range of skills in the fields of thermal engineering, residential engineering, urban engineering and mathematical engineering. Its work in these areas can be summarised by the following keywords: energy, building physics, energy systems, intelligent networks and buildings, materials, innovative processes and systems, modelling and design.

Over the last decade, the importation of fossil fuels has accounted for more than 85% of Réunion's primary energy consumption (transport, heat and electricity). The volume of importations is a major burden on the local economy and is harmful to the environment. The depletion of fossil fuels, the island's isolation and its lack of connection to the mainland's electricity grid are forcing Réunion to make an energy and ecological transition towards the use of low-consumption systems and renewable energies (both of which are environmentally friendly). The political, social and economic challenges therefore lie in energy autonomy, the efforts to tackle fuel poverty, energy management and electrical safety, while at the same time enhancing our social, ecological and human environments. The Piment laboratory is helping Réunion, France and Europe to meet these strategic challenges. It is also helping the countries of the Indian Ocean region to achieve their objectives within the framework of local, national and European policies.

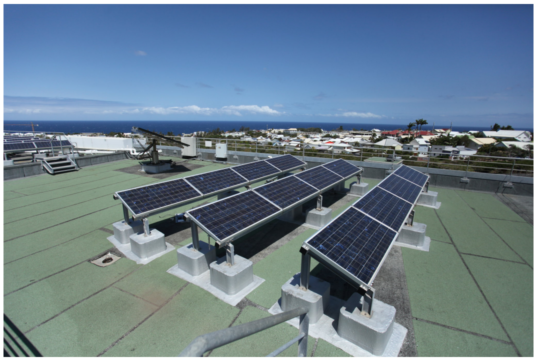


Figure 1.1: IUT Pilot site

Figure 1.2: Location of the internship

As part of its research work, Laboratoire Piment is involved in a number of actions and funding programmes at different levels: the H2020 programme, the Regional Innovation Strategy (SRI) and the Intelligent Specialisation Strategy (S3), the National Scientific Research Agency (ANR) programme, the Association of French Electrical Engineers (ADEME) programme, the GRENELLE III law, the State and Regional programmes, the Regional Climate-Energy Programme (PRCE), the Sustainable Development Programme (PDD), the Sustainable Development Programme (PDD) and the Sustainable Development Programme (PDD). the regional climate-energy programme and regional energy management.

The laboratory's most important fundamental area is generalized system dynamics. The work carried out by the laboratory ranges from "upstream" fundamental research to "downstream" fields, and it is heavily involved in collaborations in the socio-economic field. After decades of development, the laboratory has developed its own speciality. They are able to address energy issues from the micro spatial and/or temporal scale to the geographical scale.

With over 60 collaborators, PIMENT is a major player in these fields at international, national and regional level. The research carried out within the PIMENT Laboratory is generally long-term research in applied fields.

The laboratory is structured around 3 research themes,

- Energy efficiency in space and the building environment
- Sustainable energy
- Mathematics and applications

1.2 Project context

The energy transition requires the massive integration of intermittent renewable energies into electricity grids. These energies are intrinsically highly variable in space and time, and in order to control the consequences of these variabilities on grid stability and security, this integration requires the implementation of a number of resources (forecasting, storage, etc.). Many studies have highlighted the fundamental role of resource forecasting in increasing the share of solar energy in electricity grids. In this respect, probabilistic forecasts are becoming an indispensable tool, and as the penetration of these energies increases, the importance of having high-quality probabilistic forecasts becomes more and more crucial. In recent years, the scientific literature has proposed a number of probabilistic metrics for assessing the quality of this type of forecast. Examples include the CRPS, the ignorance score, the CRIGN and the Error-Spread score.

There are several freely available programming tools to help users calculate these metrics. However, the underlying assumptions are not always clearly specified, and there are inconsistencies between the various tools. In addition, there are no educational tools designed to explain clearly to non-experts the issues and good practices associated with the use of probabilistic metrics.

Faced with this situation, the PIMENT laboratory and the Observation, Impacts, Energy (O.I.E.) Centre at Mines ParisTech, as part of the International Energy Agency's PVPS programme, are proposing to develop an educational tool bringing together recommendations for calculating these metrics, simple and educational examples of calculation, and comparisons of the results obtained for different metrics on different simple examples of forecasts. The final product will take the form of an interactive teaching tool (using Jupyter notebook, for example). Finally, this work will be published in a scientific journal.

We have already mentioned that forecasting has become an important tool for the operation of intermittent renewable energy generation assets. In many cases, the quality of the forecast, i.e. the degree of statistical correspondence with observed data, is a fundamental issue for users (Le Gal La Salle 2021). Several extremely simple measures are available to assess the quality of deterministic forecasts, such as MAE, CRPS, bias, etc(Yang et al. 2020). However, probabilistic forecasts have demonstrated their value in many cases.

In short, despite the availability of tools such as CRPS, ignorance score, PINAW, etc(Lauret, David, and Pinson 2019),(Hamill 2001).And theoretical frameworks for assessing their quality (Pinson et al. 2007), they are not so simple to use.This difficulty should not be an obstacle to the adoption of probabilistic forecasts.In order to make these tools more accessible, the PIMENT laboratory at the University of La Réunion and the Université des Mines de Paris-Haute-Coeur (Sophia Antipolis), as part of the International Energy Agency's Task 16 PVPS (”Solar Energy Resources for High Penetration and Large Scale Applications”), jointly created the ”Revealing Probabilistic Metrics and their Implementation” project. The aim of the project was to create a pedagogical and educational tool, such as a website providing the code, simple examples of the use of the most common probabilistic metrics and comparisons of the results of simple probabilistic predictions.

1.3 Project objectives

CRPS is the most commonly used probability measure and a recommended methodology for calculating CRPS has been established. By following this established methodology, it is hoped that the project will refine the work on CRPS and eventually attempt to apply it to other probability measures:

- Test the calculation of these different metrics for different sets of forecasts under different assumptions,
- Compare the results obtained numerically with the reference results,
- Calculate the margins of error due to the assumptions chosen,and propose a set of calculation recommendations,
- Propose assumptions allowing a good compromise between accuracy and complexity/calculation time,
- Be able to present the results in the form of a didactic and interactive tool,
- Be able to present the results in the form of a scientific article in a peer-reviewed journal.

In addition, some specific work needs to be done, including a discussion of reliability diagrams versus CRPS reliability, a discussion of sample sizes in calculating CRPS using the Brier score (related to the equivalence of the two CRPS methods), and exploration of existing literature and tools/software for calculating CRPS in order to compare aspects of the differences and effects of existing tools and the code that will be opened up as part of the project. Assess the consistency of the description, interpretation and use of the CRPS components.

2. Context and prior concepts

2.1 Internship tasks

As part of the project as a whole, the internship includes part of the practical implementation of the project objectives and the integration of the whole project at the end, as well as the formulation of new, more specific objectives to improve the project and deepen the understanding of the relevant background knowledge.

More specifically, internship projects include the following:

- In the context of the CHPeEn predicting dataset, the effect of the number of ensembles and the integration step (i.e. threshold division) in computing CRPS when using the Brier score method is verified and computational recommendations are given.
- A summary of existing Python libraries that include CRPS, a summary of algorithmic principles and a comparison of calculations.
- Interpretation and numerical implementation of equivalence calculations for CRPS using the Quantile Score.
- Numerical implementation of the component decomposition of Quantile Score and relationship with the CRPS component decomposition
- Synthesis of analytical expressions for CRPS and quantile functions for predictions via parametric distributions that can arise in probabilistic forecasts.
- Creation of interactive tools for aggregating project tasks via jupyter book.

2.2 Introduction to the probability metric

Probabilistic forecasting plays an important role in the field of forecasting. Unlike deterministic forecasting, probabilistic forecasting is used to predict the probability of future changes in the occurrence of a relevant quantity or specific event. Probabilistic forecasts provide more complete probabilistic statistical information about the subject of the forecast, including the uncertainty of the forecast, in terms of quantity estimates, prediction interval estimates and output probability density estimates, and are therefore widely used in many fields such as meteorology, climatology, hydrology, economics and finance.

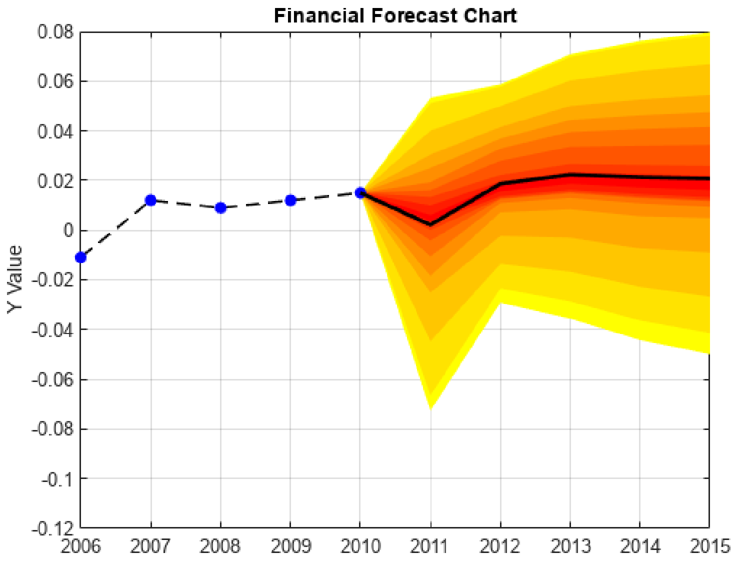


Figure 2.1: Example of probabilistic predictions

With the vigorous development of probabilistic forecasting models, the number of probabilistic models is increasing and their functions are expanding. As a result, how to effectively extract and quantify the information contained in probabilistic forecasts and how to evaluate the quality of probabilistic forecasts have become very important topics in the decision-making process. According to the existing literature, probabilistic forecasting models have been more widely used for wind energy forecasting than for solar energy forecasting, so that the field of solar energy forecasting and the related field of weather forecasting do not have enough evaluation measures to assess the quality of probabilistic forecasts. To this end, The article(Lauret, David, and Pinson 2019) proposed in 2019 a probabilistic metric framework for solar energy quality assessment.The concept of quality is proposed for the following reasons.In the field of weather forecasting, the quality of weather forecasts can be assessed in three categories, namely consistency, quality and value.

In general, people are more interested in the quality indicator, i.e. the correspondence between the predicted distribution and the observed distribution. This is why quality-based evaluations such as the CPRS score, the QS score, the Brier score, the ignorance score and many others have been proposed. These scoring rules can be expressed by the function $S(\hat{F}, x)$, and they can be interpreted as an evaluation of the accuracy of the predicted distribution $\hat{F}$ under the hypothesis that there is an observed result x in the existing distribution.

Based on the interpretation of the definition of quality, scoring rules can be seen as a measure of the error of a distribution function, for example, CRPS can be seen as a degradation of MAE (Hersbach 2000), represented by the absolute error MAE, which can be denoted by $S(x, y)$, i.e. in this case $S(\hat{F}, y)$ measures the performance of the point prediction x, whereas the CRPS notation rule, which is popular in the field of probabilistic forecasting, measures the predictive performance of the distribution F, in the more general and universal case.

2.2.1 Preliminary concepts

Proper score

The concept of proper score functions goes back to the articles of Brier 1950, Brown 1970 and Good 1952, with a more technical description in Savage 1971 and Schervish 1989.

The score function takes a probability prediction $\hat{F}$ and projects it from a set of possible probability predictions space P and an element of the sample space $y \in \Omega$ to a real value s$(\hat{F}, y) \in R$. A probability prediction can be expressed in terms of a cumulative distribution function(CDF), a probability density function(PDF), a threshold exceeding probability, or a quantitative value at a given probability level. The score function is defined in this study as the cost function that the predictor wishes to minimis , which is called negatively oriented (e.g. Gneiting and Raftery 2007).

where S$(\hat{F}, Q)$is the expected score given a specific forecast $\hat{F}$:

$$S(\hat{F}, Q) = \int_{\Omega} s(\hat{F}, y) dQ(y) \tag{2.1}$$

where the distribution of y for a fixed $\hat{F}$ is given by $Q(y) = F(y|\hat{F})$. The score function is proper if $S(Q, Q) \leq S(\hat{F}, Q)$ for all $\hat{F}$ and strictly proper if equality is given, if and only if $\hat{F} = Q$ (Gneiting and Raftery, 2007).

In summary, a scoring rule is called a "Proper" if it encourages and rewards correct predictions of true probabilities. In other words, for any possible true probability, predicting that probability should give a better expected score than predicting any other probability. Conversely, Improper scoring rules it may discourage predictors from reporting their true beliefs. In other words, this scoring rule may be a situation in which it is less preferable for predictors to report their true beliefs about the probability of an event occurring than to report other probabilities. This may reduce the quality of the predictions, and some important information may be missed. In this report, if not particularly mentioned, by default all references to Score are to the Proper Score.

Brier score

Brier Score(Brier 1950) is a commonly used proper score. It is defined as the square of the difference between the predicted probability and the actual outcome. Better predictions (i.e., predicted probabilities that are closer to the actual outcome) will result in a lower (better) Brier Score.Both our notions of Brier score and CRPS below are proposed based on the scenario of ensemble prediction, i.e., the definition is based on N observation-prediction pairs thus assessing the quality of that ensemble prediction.

$$BS = \frac{1}{N} \sum_{n=1}^{N} (\hat{F}_{1,n} - y_n)^2 \tag{2.2}$$

Quantile score

The quantile score is used to evaluate a particular type of forecast: the forecast of a quantile associated with a probability level $\tau \in [0, 1]$ defined in advance. In other words, the forecast $\hat{F}_\tau$ of probability level τ is an estimate of $Q^{-1}(\tau)$ (Where Q^{-1} stands for the quantile function here). Bentzien and Friederichs 2014 use this kind asymmetric so-called Pinball loss function to compute the error associated with a forecast/observation pair

$$QS_\tau = \tau * H(F_\tau < y) * (y - F_\tau) + (1 - \tau) * H(F_\tau > y) * (F_\tau - y) \tag{2.3}$$

Where H stands for heaviside function, (also Indicator function)

2.2.2 Review of Classical Methods for Computing CRPS

The CRPS (Continuous Ranked Probability Score) is a commonly used proper score adapted to the evaluation of forecasts of continuous variables such as the GHI. Its formulation is given by :

$$CRPS = \frac{1}{N} \sum_{n=1}^{N} CRPS_n(\hat{F}, y) = \frac{1}{N} \sum_{n=1}^{N} \int_{-\infty}^{\infty} [\hat{F}_n(x) - H(x - y_n)]^2 dx. \tag{2.4}$$

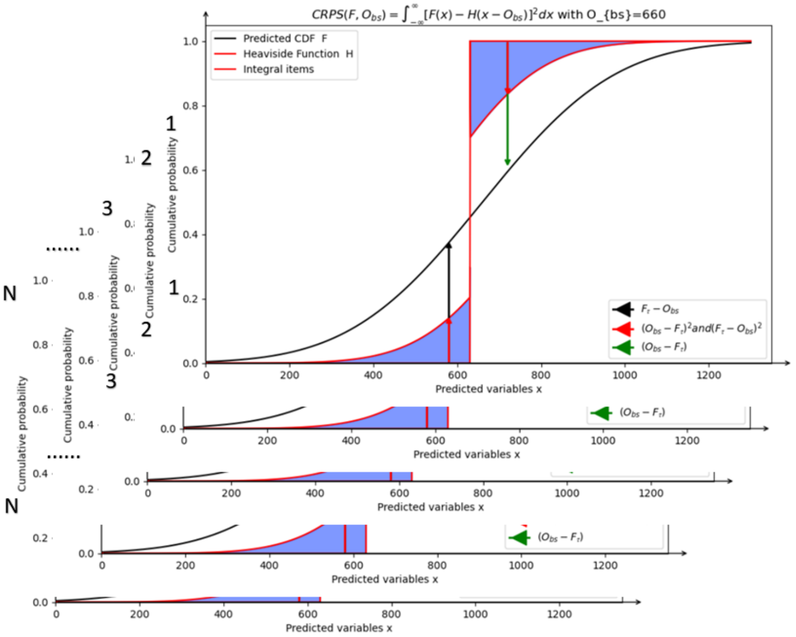


Figure 2.2: Average the scores for each observation-prediction pair

Figure 2.2 above shows the prediction of a Gaussian-type parameter distribution, but in the case of more general one,that is, the ensemble forecast, the cumulative distribution function of the probability distribution is not so easy to obtain.The common method by sorting the members of the ensemble prediction, and then converted them into Quantiles which are derived from a set of order statistics, and then the cumulative distribution function is obtained by interpolation through the quantiles. The cumulative distribution function obtained can be combined with the observations, so that the CRPS value of a forecast-observation pair can be calculated.

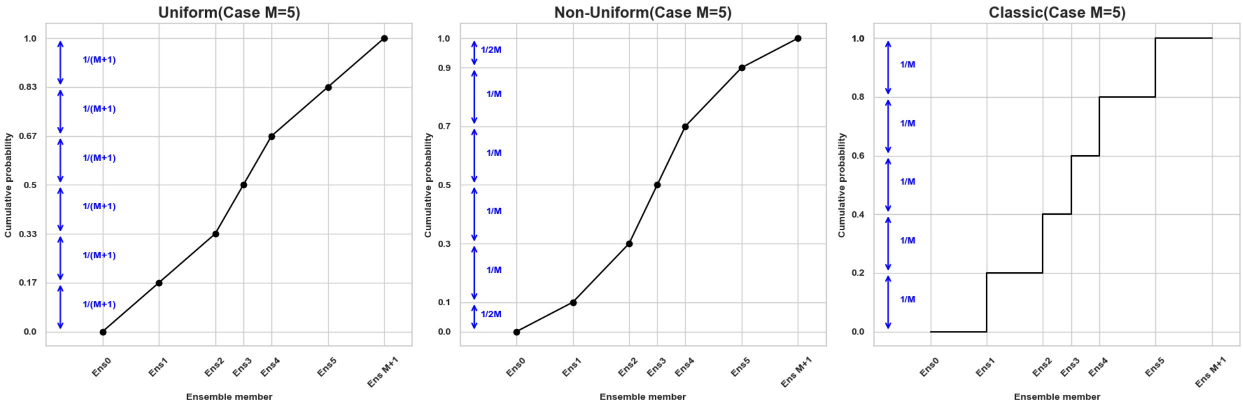


Figure 2.3: The construction method of CDF

The above three CDF construction methods are the most common three methods for modeling the probability distribution of set predictions. Among them, the classic method uses the same probability distribution for each set member, ignoring extreme cases, and CDF as a ladder function presents. The uniform method assigns probabilities to the ensemble predictions outside the range of the common probability ensemble forecast members, and the non-uniform method assigns the corresponding Ens_0, Ens_{M+1} a probability smaller than other normal probability ensemble forecast members. The other ensemble forecast members are assigned the same probability of $\frac{1}{M}$. They have an impact on the value of CRPS when using this classic definition, because the interpolation method determines the specific positions and values when we insert observations into the prediction ensemble, and the effect of this influence will be discussed in later work.

3. Scientific computation for probabilistic metrics

3.1 Introduction to Datasets

In the field of solar forecasting, probabilistic solar irradiance forecasts are often benchmarked using the Clear Sky Afterglow Ensemble (PeEn). The skill score is obtained by comparing the continuous rank probability score (CRPS) of the predictive model with the continuous rank probability score of PeEn. But the complete-history PeEn (CH-PeEn) proposed by Yang 2019's paper,which is more general as a benchmark method for probabilistic solar forecasting. CH-PeEn utilizes the entire measurement history and forms an empirical distribution of the predicted clear sky index that depends only on the time of day, so the ensemble-forecast data in this report will be generated under the framework of CH-PeEn for each of the eight stations.The assumptions, analyzes and variation phenomena of the scores related to the observations and predictions of the eight sites are discussed below.The geographical information and abbreviated identification of the sites are presented in Table 3.1.

Site	Abbreviations	Latitude	Longitude	Altitude
Cabauw	CAB	51.9711	4.9267	0
Carpentras	CAR	44.083	5.059	100
Cener	CEN	42.816	-1.601	471
Milan	MIL	45.47618	9.254559	150
Palaiseau	PAL	48.713	2.208	156
Payerne	PAY	46.815	6.944	491
PlataformaSolar	TAB	37.0909	-2.3581	500
Toravere	TOR	58.254	26.462	70

Table 3.1: Geographic information and abbreviations for dataset sites

3.2 Overview of relevant Python libraries

As we introduced in the background at the beginning of the report, for the calculation of CRPS, there are many R language software packages(scoringRules,verification,ensembleMOS,etc.) and Python packages(Xskillscore,CRPS,Properscroing,etc.) that introduce related functions to help climate prediction researchers and related application industries to calculate CRPS of data, but to us, they are more or less flawed. It can be mainly summarized as the following points:

1. No component decomposition for the general proper score.

   For common scores, Brier scores and Quantile scores, and CRPS have only numerical results, ignoring the description and decomposition calculation of reliability, resolution and uncertainty.

   That is, through the mapping of the score function, a single value in the real number domain has been taken as an evaluation reference. For graphical representations in relevant fields like weather forecasting and solar energy prediction, there are no corresponding references to aid in understanding for the reliability diagrams and rank diagrams. Furthermore, the uncertainties brought by sampling of the original observational data have not been accurately quantified.

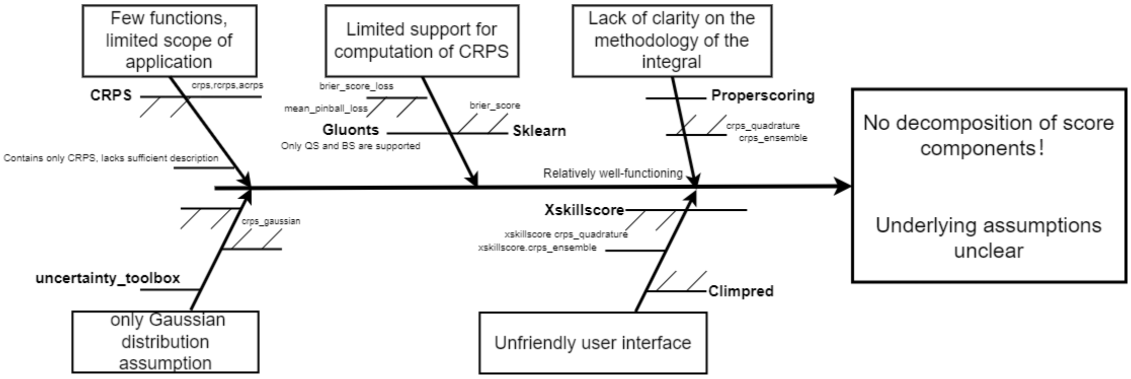

Few functions, limited scope of application
Limited support for computation of CRPS
Lack of clarity on the methodology of the integral
CRPS
crps,rcrps,acrps
Contains only CRPS, lacks sufficient description
brier_score_loss
mean_pinball_loss
Gluonts
Only QS and BS are supported
brier_score
Sklearn
Properscoring
crps_quadrature
crps_ensemble
Relatively well-functioning
No decomposition of score components!
Underlying assumptions unclear
crps_gaussian
uncertainty_toolbox
only Gaussian distribution assumption
Xskillscore
xskillscore crps_quadrature
xskillscore.crps_ensemble
Climpred
Unfriendly user interface

Figure 3.1: Python library for calculating CRPS

This results in the score available for the decision-making process being neither comprehensive nor precise enough.

2. The underlying assumptions are unclear.

   For example,the computational functions of Climpred, Xskillscore and Properscroing are all created based on the original functions of Properscroing, and the functions to compute ensemble predictions tells us that the weights parameter is equal by default.

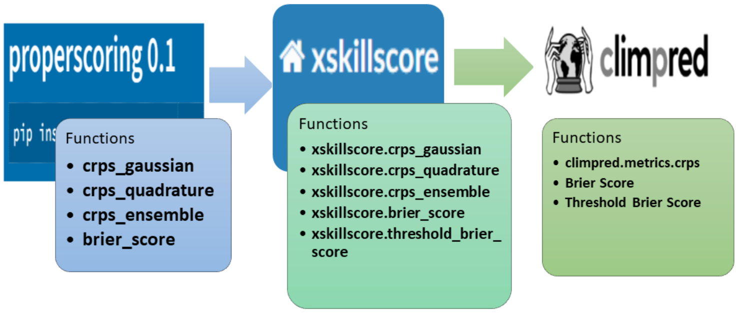


Figure 3.2: Function profiles of three typical libraries for computing CRPS

According to the description of the relevant references (Hersbach 2000), the weights may have different weights according to different grid points or classifications. Specifically, the weights can be proportional to the cosine of the geographic location's latitude, which is common in the context of geosciences or meteorology, as this takes into account the influence of the geometry of the sphere on the calculation. But in practical applications, when our data source is based on a time series prediction of a certain location, this weight is easily misunderstood as the weight of the time point. Although we can certainly apply weight to the time point, but we cannot directly pass it to the function to accomplish. In this case, the weight parameter can only be interpreted as the probability weight assigned to each ensemble forecast member when constructing the probability cumulative distribution function (CDF) of the ensemble forecast.However,in this case it differs from its original meaning

Correspondingly, the document only provides a simple calculation example, ignoring how to construct the probability representation of the set prediction, that is, express the prediction result through the process of obtaining CDF, PDF and quantile, and only includes raw data calculation examples, which cannot help non- Professionals understand the subject well and implement the application.

3. Methods that are not directly applicable to non-experts.

   Some functions of these libraries need to convert to a specific data format, or apply a cyclic format when working with large datasets, so the user interface is not very good. For example, this requires us to either construct a class and use the attribute calculation function (Xskillscore,Climpred,etc.), or convert the xarray type, or loop through the dataset (Properscoring).

4. Some libraries only calculate CRPS by posing Gaussian distribution assumptions.

   It is only assumed that the probability predictions are Gaussian, which is only applicable to specific problems in special contexts, and predictions and evaluations are made at the same time.Thus, the scope of application of the related library for the calculation of scores has not been expanded.Although this kind of prediction has no error and estimation of numerical calculation, the CRPS value is directly given by analytical expression, and there is no general modeling method for collective prediction, so the scope of application is relatively narrow.Only from the perspective of parameter distribution, there are many other predictions of parameter distribution that can be applied to different prediction scenarios.For example, forecasting rainfall through gamma distribution, reliability analysis and survival analysis through Weibull distribution,etc.

Correspondingly, the functions of the Python calculation CRPS library also have advantages, which are mainly reflected in the calculation speed and support for multidimensional arrays. This optimizes the time complexity of the calculation, avoids the possibility of calculating the predicted observation pairs at different time points in the loop structure, and directly operates the matrix to make the calculation more efficient. We will illustrate this later in the computational implementation section.[1]

Site	*ps.crps_ensemble*(time/s)	*ps.threshold_brier_score*(time/s)	(*sklearn*)*mean_pinball_loss*(time/s)
CAR	78.88(1.16)	78.91(1.05)	78.91(28.21)
CAB	89.04(0.96)	89.04(1.05)	89.07(25.32)
CEN	86.02(1.34)	86.02(0.94)	86.05(31.98)
MIL	86.20(1.02)	86.30(1.12)	86.24(24.54)
PAL	88.36(0.92)	88.35(1.35)	88.40(23.45)
PAY	93.79(0.94)	93.85(1.11)	93.84(23.26)
TAB	62.73(1.27)	93.85(1.28)	62.73(26.87)
TOR	78.85(0.48)	78.60(0.61)	78.88(26.87)

Table 3.2: The time for three methods for calculating CRPS using existing libraries

In the table, the CRPS calculated by the classical method (*ps.crps_ensemble*) and the Brier score method using the existing library is completely consistent with the CRPS result obtained by our custom function(Which takes eight to ten times more).

Thus, it also shows the reasonableness of our next explanation of the hypothesis and calculation.However, as we mentioned, the vectorized (*ps.crps_ensemble*) function has more advantages in terms of computational

[1] ps in the table 4.2 refers to Python library,Properscroing

efficiency. If you do not consider the method of constructing the CDF and the problem of score decomposition, such as uniform and non-uniform, which will be mentioned later, for the set under the classical definition Prediction calculations are more efficient.

In addition, using quantile score to calculate CRPS is not practical in terms of calculation cost. From this perspective, more attention should be paid to considering the meaning of quantile score itself and its decomposition components.In the appendix, there are more details to summarize the assumptions and approxiamation of the function in Python libraries(including Properscoring,CRPS,etc.).

3.3 Computational recommendations for classical and Brier score methods

3.3.1 Computation of CRPS under classical definition

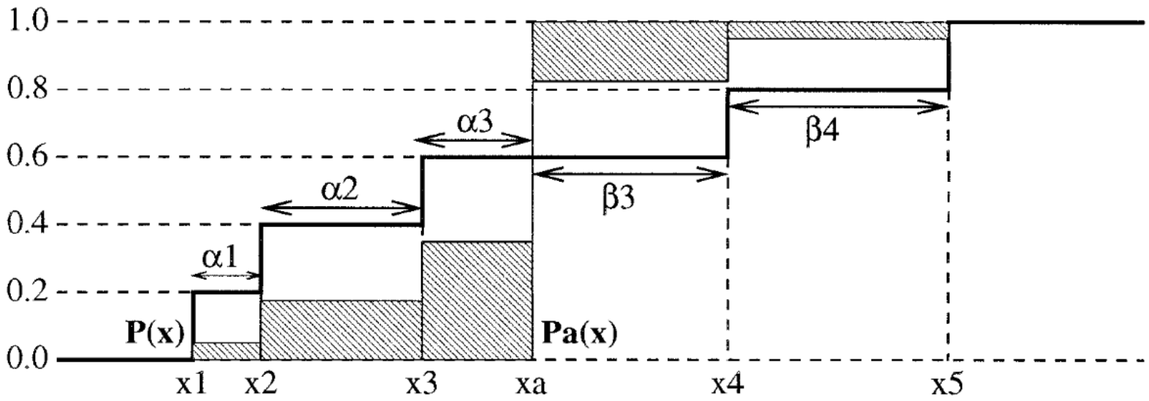


Figure 3.3: Calculation of CRPS of ensemble forecast(Adapted from Hersbach 2000)

Here a five-member set x1,...,x5 is taken as an example (thick solid line) and the cumulative distribution of the verification analysis xa (thin solid line), CRPS is indicated by the shaded area. α_i and β_i each represent the width between predicted values of ensemble members before and after exceeding the validation observations.

Here is an example of the CDF constructed by the classical method. In fact, for the classical method, after inserting the observed value into the CDF, it corresponds to the interp1d(fc, prob, kind='previous') interpolation function in python, that is, the total Returns the value of the CDF corresponding to the previous ensemble member.

For the CRPS determined by this type of CDF, when we do discrete integration, we naturally think of rectangular integration. But it is worth noting that for the CDF constructed by the classical method, the discrete integral of the left rectangle should be used. Although the experimental results show that the values of the discrete integrals of the left and right rectangles are not much different, if the systematic error is used, the integral term of the CRPS is always overestimated for the first half of the CRPS because of the jump in the right endpoint value (that is, the part of the ensemble member prediction that does not exceed the validation observations). For the other half, due to the minus sign, the integral term of CRPS is underestimated, and we cannot evaluate whether the overall shaded area is overestimated or underestimated, because it obviously depends on the variation law of the set,$\{\alpha_i\}$ and $\{\beta_i\}$ values. If using the discrete integral of the left rectangle, no matter how and where the verification value is inserted, this error can be avoided.

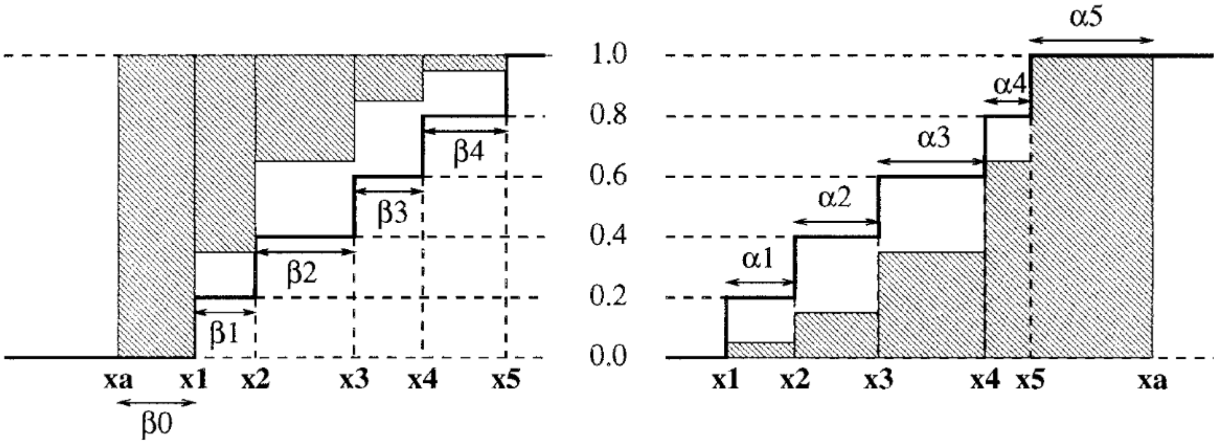


Figure 3.4: Outlier case for calculation of CRPS of ensemble forecast(Adapted from Hersbach 2000)

The aforementioned discussion still applies to situations where the validation values fall outside of the predicted ensemble.

3.3.2 Verification for the stability of the threshold integration step

Using the left integral method under the classical CDF framework to obtain a CRPS value with no error numerical estimation.Hersbach 2000 shows that CRPS can be computed by integrating the Brier scores over all possible predictors because after we replace the predictor variables with threshold variables, it is not difficult to see that $\hat{F}_{1,n} = \hat{F}_n(x)$ and $y_n = H(x - y_n)$. Mentioned in Section 2.2.1 that Brier Score (BS) is a scoring rule used to predict the occurrence of a specific event. Typically, such events are characterized by a threshold x. The event occurs if $x < y_{obs}$ and does not occur if $x > y_{obs}$. Then it can be derived,

$$CRPS = \int BS(x)dx \tag{3.1}$$

Therefore, we divided different thresholds equidistantly, and took the distance between the two thresholds as the integration step. it is easy to be found that when the distance between the GHI thresholds is less than $50W \cdot m^{-2}$, taking the result of classic definition of CRPS as a standard reference, the error of CRPS estimated by Brier score has already stabilized, which demonstrates the equivalence of the two methods numerically.Moreover,for the area of estimating the quality of solar forecasts (with GHI as the predictor variable), it can be infered that the CRPS obtained through the Brier score is credible when the distance between the GHI thresholds is less than $50W \cdot m^{-2}$.

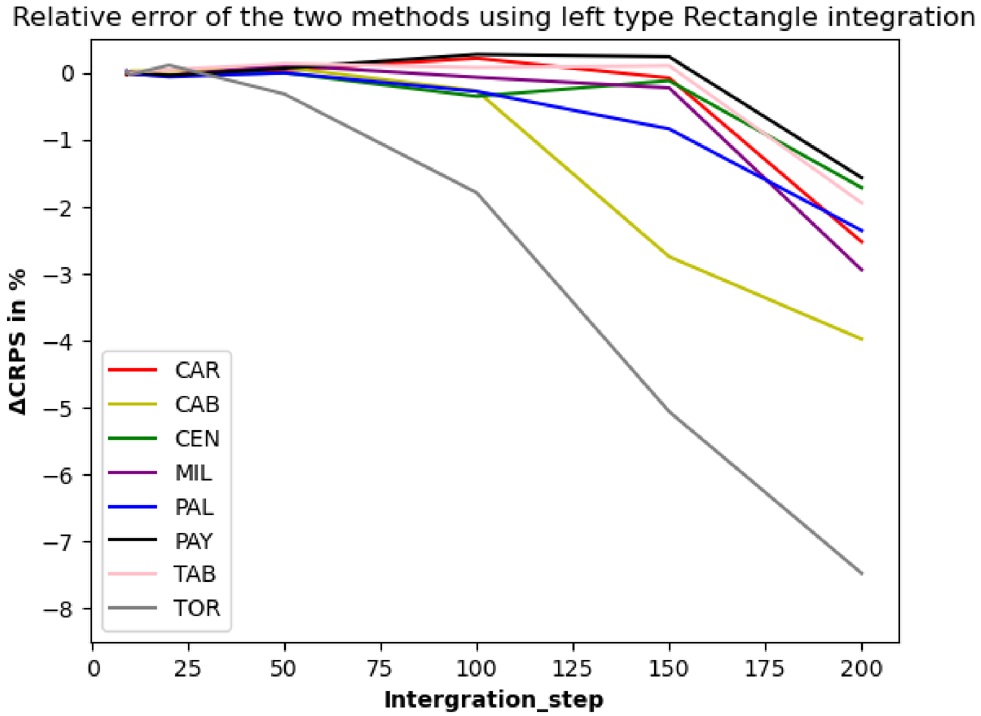


Figure 3.5: The effect of Brier score integration threshold

3.3.3 The Disscussion of Different CDFs on Integral Methods

As discussed in the previous section, under the framework of the classical method of constructing CDF, we should choose the integration method of the left rectangle to obtain the accurate CRPS value under this method. What if the CDF construction method becomes a uniform method or a non-uniform method ?

In this case, since the connection between the two ensemble members on the image becomes linear instead of returning the value of the previous ensemble member, when inserting the verification observation value, the CDF truncation point corresponding to the verification observation value, which needs to use the linear function interp1d(fc, prob, kind='linear')to interpolate.To confirm that when the original linear CDF is used as a part of the integral item for calculating CRPS, where it is represented by p^2 and $(1-p)^2$ respectively, so the corresponding value of CRPS is Figure 3.6 The area of the shaded part under the corresponding curve , each segment of these shaded parts (that is, the connection between every two ensemble members) is connected by a quadratic curve, so we want to get the area of the shadow part, using the Simpson integral method can get the numerically accurate value of CRPS under the framework of constructing CDF under the uniform and non-uniform method.

For both uniform and non-uniform methods, climatological extremes are usually chosen to be assigned to the corresponding Ens_0 and Ens_{M+1} in Figure 3.6 In this report, a value of $-4W \cdot m^{-2}$ is chosen to assign a probability of 0 to Ens_0, and Ens_{M+1} of $1300W \cdot m^{-2}$ is the empirical climatological maximum.This choice is because Long and Shi (2008) propose $Clim_{min} = -4W \cdot m^{-2}$ and $Clim_{max} = 1.5 \times S_a \times \mu_0^{1.2} + 100W \cdot m^{-2}$ where S_a is the solar constant and μ_0 is the cosine of the zenital solar angle.

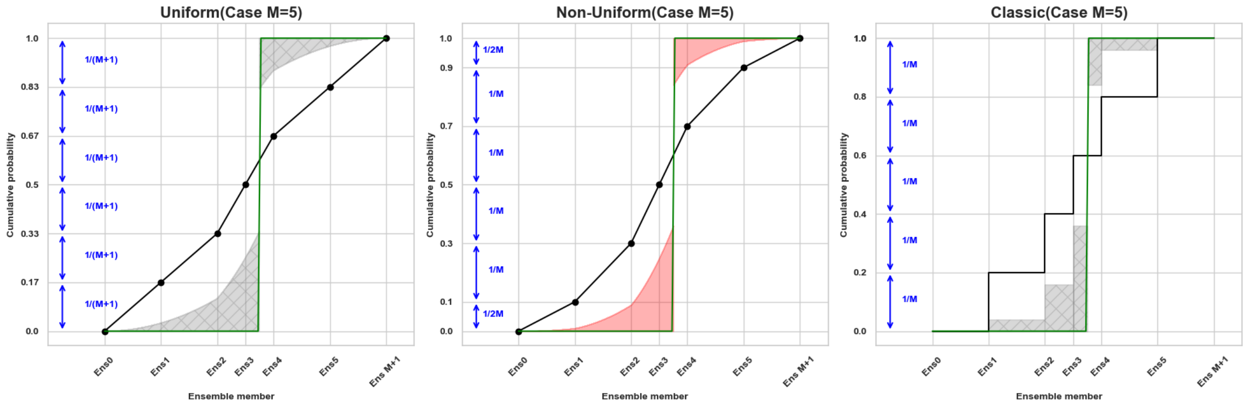


Figure 3.6: Intergration with different construction method of CDF

The specific practical method steps are as follows,

- Confirm the cumulative probability value of the midpoint of two ensemble members by linear interpolation
- Two endpoints and a midpoint uniquely define a quadratic curve
- Use the Simpson integral formula(Simpson's 1/3 rule)

$$\int_a^b f(x)\,dx \approx \frac{h}{3}\left[f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right] \tag{3.2}$$

where $h = \frac{b-a}{2}$. More generally, if we integrate the entire interval, Simpson's 1 rule can be expressed as, because each sub-interval is divided into two small sub-intervals by taking the midpoint, n is the number of sub-intervals, so where n must be an even number.

$$\int_a^b f(x)\,dx \approx \frac{h}{3}\left[f(a) + 4\sum_{i=1}^{n/2} f(x_{2i-1}) + 2\sum_{i=1}^{n/2-1} f(x_{2i}) + f(b)\right] \tag{3.3}$$

where $h = \frac{b-a}{n}, and \quad x_i = a + ih$ Obviously, these quadratic curves are in one-to-one correspondence with the quadratic curves enclosed in the shaded part of the actual image.

Method of CDF construction	Integration method
Classic	Left Rectangular Integral
Uniform	Simpson's 1/3 Integral
Non-uniform	Simpson's 1/3 Integral

Table 3.3: Summary of integration methods corresponding to different CDFs

Theoretically, these integration methods will achieve no discrete error integration for the case where the observed value is outside the ensemble prediction, inside the ensemble prediction, and between two members of the ensemble prediction.

In fact, for the method of constructing CDF for higher-order curves, due to the consideration of computational cost and the actual result of CRPS (the value of CRPS is not much different under the premise of constructing CDF for different higher-order curves) , it is not very necessary to find an integration no-error method under the other infrequently used CDF construction methods. When controlling the desired integration accuracy within a certain range,some adaptive integration methods (such as the quad function in scipy.integrate) can meet the needs well.In addition, within a certain range of error tolerance, we can also consider using the trapezoidal method instead of the Simpson method to integrate the uniform and non-uniform linear CDF methods.

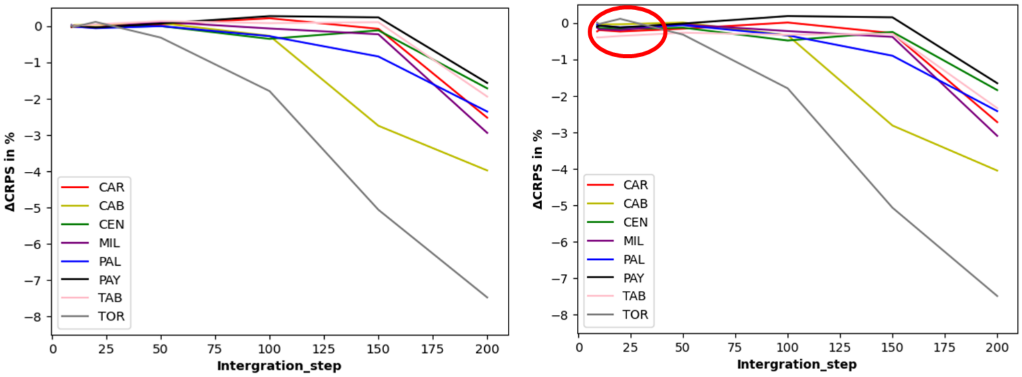


Figure 3.7: Relative errors of two methods using the Simpson(Left) and Trapezoidal(Right) method

From Figures 3.7 and error trends in red circles of the picture on the right(Trapezoidal method), it can be seen that for investigating the accuracy and stability of the Brier score threshold division of the integration step for integrating to obtain the value of CRPS, choosing the appropriate discrete integration method without error helps us to make the appropriate judgment because the integration method determines the classical definition CRPS and we want to choose as the reference standard/benchmark. Numerically, it should have consistency. Although the numerical difference between the trapezoidal integration method and the Simpson's integration method is very small(Refer to the table 3.4), it is more accurate to choose the Simpson's integration method as a reference for giving recommendations on the stability range.

Site	Abbreviations	Trapezoidal	Left rectangular	Right rectangular	Simpson
Carpentras	CAR	89.05990411	88.99819544	89.12161278	89.00823072
Cabauw	CAB	79.01088412	78.84933086	79.17243739	78.95791759
Cener	CEN	86.11093867	85.99835336	86.22352398	86.06730264
Milan	MIL	86.30354508	86.16880125	86.43828890	86.24680143
Palaiseau	PAL	88.36706797	88.31236612	88.42176982	88.31471203
Payerne	PAY	93.83146937	93.75100626	93.91193249	93.7861292
PlataformaSolar	TAB	62.97748106	62.72346282	63.23149931	62.89536988
Toravere	TOR	78.82367146	78.82156705	78.82577587	78.79161996

Table 3.4: Results for different integration methods under the classical CRPS definition

However, many existing python libraries directly call the quad function as the integration method of the ensemble forecast. If there is no prerequisite for CDF construction, it is difficult for users to know how to choose the margin of error for different intergration method with different CDF.

3.3.4 Decomposition of Brier Score and CRPS

The article of Bröcker 2008a pointed out that for proper score, it can be considered that there is a general decomposition method. By introducing concepts[②] such as divergence and entropie of the distribution Ω for average score($\frac{1}{N}\sum_{n=1}^{N} s_n(\hat{F}_n, y_n)$), theoretically, we can express any proper score as in the form of

$$S(\hat{F}, Q) = REL - RES + UNC \tag{3.4}$$

Naturally, for the proper score CRPS and Brier Score, respectively,we have,

$$CRPS = REL_{crps} - RES_{crps} + UNC_{crps} \tag{3.5}$$

$$BS = REL_{bs} - RES_{bs} + UNC_{bs} \tag{3.6}$$

$$\int BS(x)dx = \int REL_{bs}(x)dx - \int RES_{bs}(x)dx + \int UNC_{bs}(x)dx \tag{3.7}$$

For each level of GHI(corresponds to the x in our integration variable), one can acquire the reliability, resolution and uncertainty components of the Brier score. The common question is if we can represent the components of CRPS by integrating the components of BS with respect to the threshold variable GHI, but it should be noted that Hersbach 2000 provides an alternative decomposition of CRPS so that linking equation 3.5 and equation 3.7 is possible through equation 3.1 from the previous subsection.

However, for this variant, the reliability and resolution differ by a constant D compared to the above classical decomposition, provided that the calculation of uncertainty is constant. It can be seen that this alternative decomposition does not preserve the interesting property that allows simple linking of Brier scores and CRPS.

$$REL_{crps} = \int REL_{bs}(x)dx + D \tag{3.8}$$

$$RES_{crps} = \int RES_{bs}(x)dx + D \tag{3.9}$$

$$UNC_{crps} = \int UNC_{bs}(x)dx \tag{3.10}$$

In general the integration of the resolution and reliability of the Brier score at all possible thresholds, respectively, generally differs from that of the CRPS.Nevertheless, the integral of all uncertainties $UNC_{bs}(x)$ is equal to the following uncertainty: UNC_{crps}. So it is still meaningful to realize the decomposition of CRPS by studying the integral on the possible threshold of Brier score.

Additionally,this part of $\int REL_{bs}(x)dx$ is stricter than the corresponding part of REL_{crps}, $\int REL_{bs}(x)dx$ insists on perfect reliability of all possible events, while REL_{crps} is more focused on the overall reliability of the forecast system.Therefore, by decomposing and integrating the Brier Score to obtain the corresponding components of CRPS, it is still possible to compare and judge the performance of multiple prediction systems by comparing the components.

[②] The details of related concepts can be found in Bröcker 2008a and Le Gal La Salle 2021, which will not be expanded here due to space reasons.

Brief description of the decomposition principle

We have already introduced the form of the Brier score through Equation 2.2, which is defined as follows:

$$BS = \frac{1}{N}\sum_{n=1}^{N}\left(\hat{F}_{1,n} - y_n\right)^2$$

with

$$REL_{bs}(x) = \sum_{k=0}^{M} g_k(x)\left[o_k(x) - p_k\right]^2$$
$$RES_{bs}(x) = \sum_{k=0}^{M} g_k(x)\left[o_k(x) - o(x)\right]^2$$
$$U_{bs}(x) = o(x)[1 - o(x)]$$

where M is the number of ensemble members,$g_k(x)$ is the discretisation weight, $o_k(x)$ is the proportion of the number of events in the corresponding discretisation group, $o(x)$ is the climatological mean, and p_k is the corresponding predicted probability value.Specific details can be found in the appendix or the paper of Hersbach 2000.

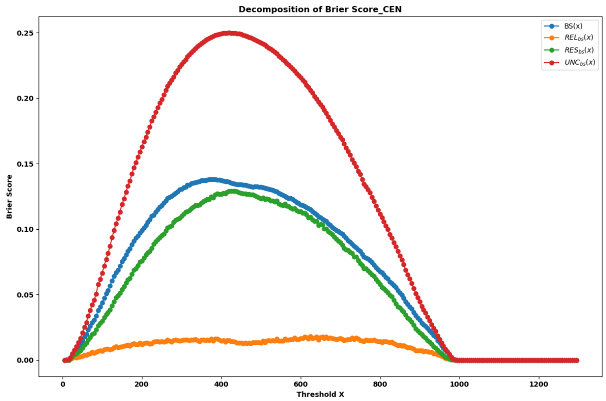


Figure 3.8: Decomposition of Brier Score(Site CEN)

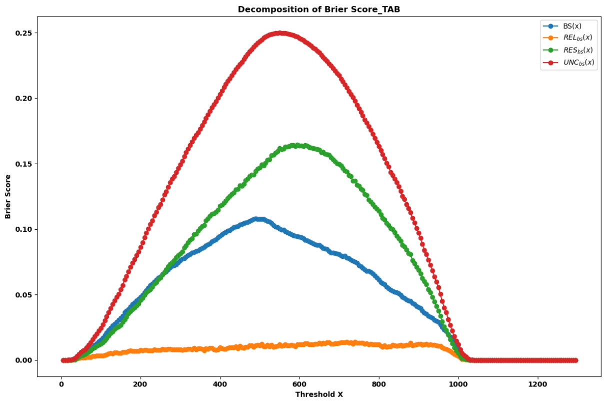


Figure 3.9: Decomposition of Brier Score(Site TAB)

The pictures from figure 3.8-figure 3.11 represents the CRPS components obtained by Brier Score decomposition, and the area under each curve corresponds to the associated CRPS component. CRPS is obtained by integrating $BS(x)$ over all thresholds x.Also worth noting, In our case, the integration over x of the different components ranges for values of GHI from 0 to the maximum of the climatetology(In the code for decomposing and integrating Brier, this value is set to $1300W \cdot m^{-2}$).

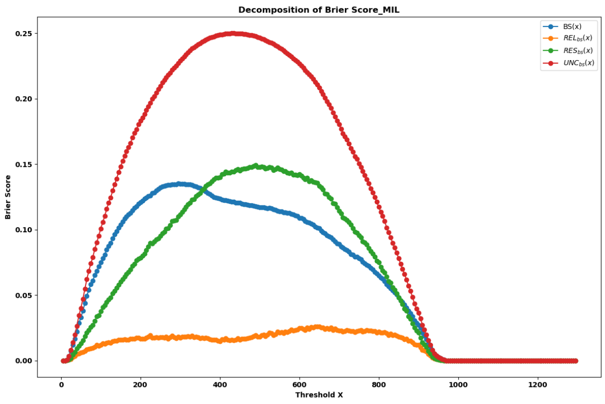


Figure 3.10: Decomposition of Brier Score(Site MIL)

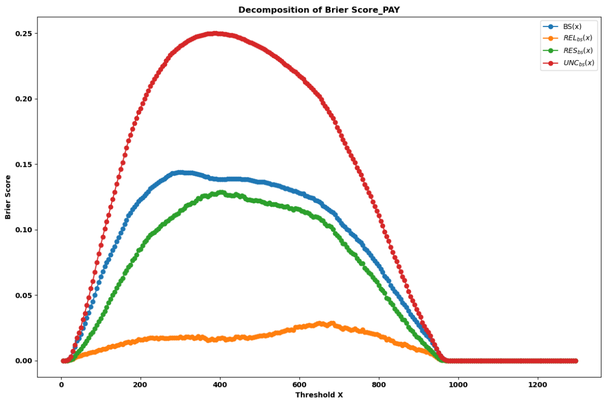


Figure 3.11: Decomposition of Brier Score(Site PAY)

3.4 Decomposition of QS scores and association with CRPS

3.4.1 Description

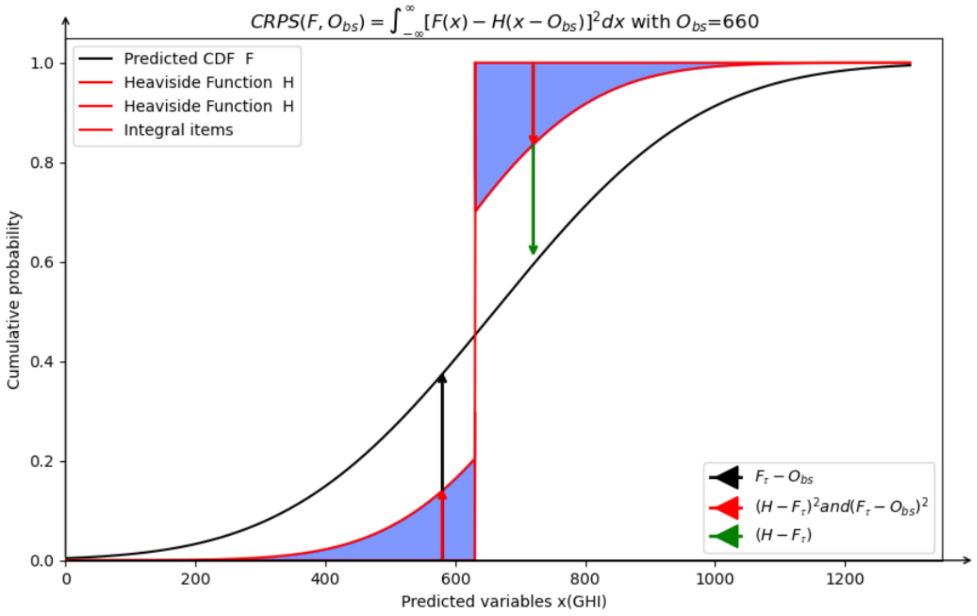


Figure 3.12: CRPS under the classical definition(N=1)

Figure 3.12 illustrates the classical definition of the CRPS as an integral against a predictor variable (which in the context of the report is often treated as the GHI), with the area of the blue shaded area equal to the value of the CRPS. Figure 3.13 puts the two methods together and compares them. The CRPS is calculated by Quantile score by integrating against the probability level of the set of predictors, so the area of the blue portion made up of the two corresponding arcs on the horizontal axis is the value of the CRPS, and the numerical integration method verifies that the two portions are equal.The next section will centre around this fact.

The Quantile score(QS) was mentioned in the section on a priori concepts in Section 2.2.1, and it was also mentioned in the Laio and Tamea 2006 article that the CRPS can be obtained by integrating a cost-loss function with respect to probability levels.

$$CRPS(\hat{F}, y) = 2\int_0^1 QS_\tau d\tau \tag{3.11}$$

where $QS_\tau = \tau * H(F_\tau < y) * (y - F_\tau) + (1 - \tau) * H(F_\tau > y) * (F_\tau - y)$[3]

Migrating this cost-loss function corresponds to the evaluation metric of probabilistic prediction, which is Quantile socre, however in the article of Laio and Tamea 2006 does not give some details, so the rederivation from the perspective of the proper score is demonstrated in the appendix.

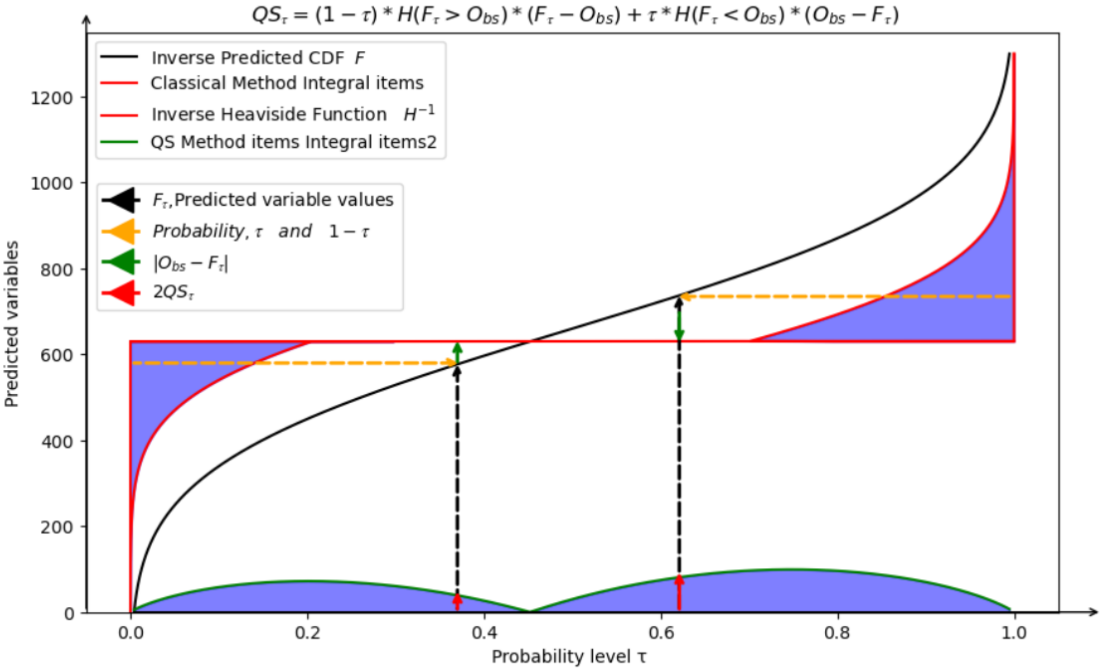


Figure 3.13: CRPS calculated by QS (N=1)

In Figure 3.11, the heights of the two arcs lying on the probability level of the horizontal axis represent two times the value of the QS score at different probability levels. This probabilistic prediction is made with

[3] O_{bs} in Figure 3.10 and 3.11 is equivalent to y in Eq 3.11

one observation as 630,And the prediction mean is set to 660 while the standard deviation is set to 250 as a Gaussian type distribution, so we see that roughly at a probability level equal to 0.45, this is a perfect prediction corresponding to a minimum Quantile score of 0. (Because the Quantile score is a negatively oriented score), and similarly for the Extreme events, according to climatology, are set to 0 and 1300, so the corresponding probability levels for anomalies smaller than that 0 and larger than 1300 correspond to 0 and 1, respectively, similar to the settings in our uniform CDF method.

In sum, the quantile score is good at capturing the predictive information of the quantile for a given probability level.Especially for the tails of the distribution, thus showing the prediction accuracy of the distribution for the quantiles corresponding to higher and lower probability levels, and also determining who is more affected by the lowest and highest quantiles of a distribution, and how different sites of the same predictive model are affected by the lower and higher quantiles.

Thus, we can obtain a CRPS by integrating over the predictor variables under the classical definition, and this CRPS is a macroscopic evaluation of all predicted ensemble members of the whole system. Similarly, for the same system, we can first evaluate individual prediction ensemble members by QS score to observe the quality of prediction at a particular probability level, and finally when we integrate on the probability level axis by replacing the integrating variables, we can also get an overall evaluation of that prediction system. This reveals the inherent equivalence between using quantiles to represent the probability distribution of a prediction system and using a CDF to represent a probability system, but in specific cases, using quantiles may be able to tap into more information about the edges and tails of the prediction distribution.

3.4.2 The principle of QS decomposition

As a proper score, the Quantile score also has its component decomposition, and it was natural to want to explore whether some information about the CRPS components could be obtained from the components of the Quantile score, as was the case in the Brier score panel.Noticing **Bentzien and Friederichs 2014** proposed to decompose QS into three components: reliability, resolution and uncertainty.

In the previous section it was mentioned roughly that all the proper scores can be represented as a sum of the three decomposed components, to introduce the decomposition of the Quantile score, a more detailed description is given below. The decomposition of the expected score is now

$$S(\hat{F}, Q) = S(\bar{Q}, Q) - d(\bar{Q}, Q) + d(\hat{F}, Q)$$

The first two items, which we usually refer to as entropy, can be expressed as $Entropy(Q) = S(\bar{Q}, Q) - d(\bar{Q}, Q)$, where $\bar{Q}$ represents the marginal distribution of the observed data, often referred to as climatology. The resolution of the forecast P is expressed as the difference between the climatological forecast $\bar{Q}$ and Q.

Similarly,the global score of the probabilistic prediction scheme is obtained by integrating the marginal distribution dF($\hat{F}$) :

$$\int_P S(\hat{F}, Q) dF(\hat{F}) = \underbrace{S(\bar{Q}, \bar{Q})}_{uncertainty} - \underbrace{\int_P d(\bar{Q}, Q) dF(\hat{F})}_{resolution} + \underbrace{\int_P d(\hat{F}, Q) dF(\hat{F})}_{reliability} \tag{3.12}$$

Additionally the uncertainty comes entirely from the characterisation of the validation data, so changes in the probabilistic forecast will only affect the reliability and resolution components of the scoring, not the uncertainty.

Given a level of probability τ, the decomposition in Eq. (3.12) for the QS is thus expressed as

$$Eq.(3.12) = \underbrace{\int_{-\infty}^{\infty} \tau(y - \bar{q_\tau})dQ - \int_{-\infty}^{\bar{q_\tau}} (y - \bar{q_\tau})dQ}_{uncertainty} - \underbrace{\int_{q_\tau}^{\bar{q_\tau}} (\bar{q_\tau} - y)dQ}_{resolution} + \underbrace{\int_{q_\tau}^{p_\tau} (p_\tau - y)dQ}_{reliability} \qquad (3.13)$$

where q_τ is the τ-quantile of the distribution Q,the $\bar{q_\tau}$ is the τ-quantile of the climatology $\bar{Q}$,p_τ is the quantile forecasts,y is an element of the sample space Ω,and $Q(y) = F(y|\hat{F})$.

Applying it to discrete ensemble prediction, consider a validation dataset consisting of $n = 1, ..., N$ a validation dataset consisting of pairs of predictions $p_{\tau,n}$ and observationso_n. The average QS is

$$QS_\tau(\hat{F}, y)_N = \frac{1}{N} \sum_{n=1}^{N} CL_\tau(o_n - p_{\tau,n}) \qquad (3.14)$$

Identical to the definition in equation 2.3, where

$$\mathrm{CL}_\tau(\nu) = \nu(\tau - H_{[v<0]}) = \begin{cases} \nu\tau & if \quad \nu \geq 0 \\ \nu(\tau - 1) & if \quad \nu < 0 \end{cases}$$

In order to estimate reliability and resolution, it is necessary to condition the observations on the predicted values. This requires rearranging and grouping (categorising) successive predicted values in order to estimate the value of the conditional measure for each category of observations.

For this purpose, we classified the data into k = 1, ... , K bins,Notated here as I_k. Bins are defined in such a way that they represent similar predictions. Assume that N_k is the number of pairs of data in each bin, $\sum_{k=1}^{K} N_k = N$. The entropy $S_\tau(Q, Q)$of each bins is estimated respectively as

$$S_\tau(Q, Q)^{(k)} = \tfrac{1}{N_k} \textstyle\sum\limits_{n \in I_k} CL_\tau(o_n - o_\tau^{(k)})$$

where the conditional quantiles are estimated as the τ-quantile $o_\tau^{(k)}$ of all observations on with $n \in I_k$.

For ensemble prediction, after we discretise the continuous predictor variables, the reliability and resolution of the QS are estimated as the difference between the mean scores of $o_\tau^{(k)}$ and $\bar{o}$, respectively,

$$d_\tau(\bar{Q}, Q)^{(k)} = S_\tau(\bar{Q}, Q)^{(k)} - S_\tau(Q, Q)^{(k)} = \tfrac{1}{N_k} \textstyle\sum\limits_{n \in I_k} [CL_\tau(o_n - \bar{o_\tau}) - CL_\tau(o_n - o_\tau^{(k)})]$$

$$d_\tau(\hat{F}, Q)^{(k)} = S_\tau(\hat{F}, Q)^{(k)} - S_\tau(Q, Q)^{(k)} = \tfrac{1}{N_k} \textstyle\sum\limits_{n \in I_k} [CL_\tau(o_n - p_\tau^{(k)}) - CL_\tau(o_n - o_\tau^{(k)})]$$

where the $p_\tau^{(k)}$ represent the discretized quantile,the average value $p_\tau^{(k)}$ of all forecasts $p_{\tau,n}$ with $n \in I_k$.And $\bar{o_\tau}$, which is given as the τ-quantile of the N observations.forecasts[4]

The decomposition Eq(.3.12) of the average QS of the discretized forecasts$p_\tau^{(k)}$, $k = 1, ..., K$, is obtained by summation over all bins:

$$\sum_{k=1}^{K} \frac{N_k}{N} S_\tau(\hat{F}, Q)^{(k)} = \underbrace{\frac{N_k}{N} S_\tau(\bar{Q}, Q)^{(k)}}_{uncertainty} - \underbrace{\sum_{k=1}^{K} \frac{N_k}{N} d_\tau(\bar{Q}, Q)^{(k)}}_{resolution} + \underbrace{\sum_{k=1}^{K} \frac{N_k}{N} \mathrm{d}_\tau(\hat{F}, Q)^{(k)}}_{reliability}. \qquad (3.15)$$

[4] For the Brier score, the representing forecast value is often set to the mid-value of the interval.Particularly if the forecast values are not uniformly distributed, the estimates of the reliability component are largely biased (see for instance Bentzien and Friederichs 2014).

3.4.3 Methodology and decomposition bias

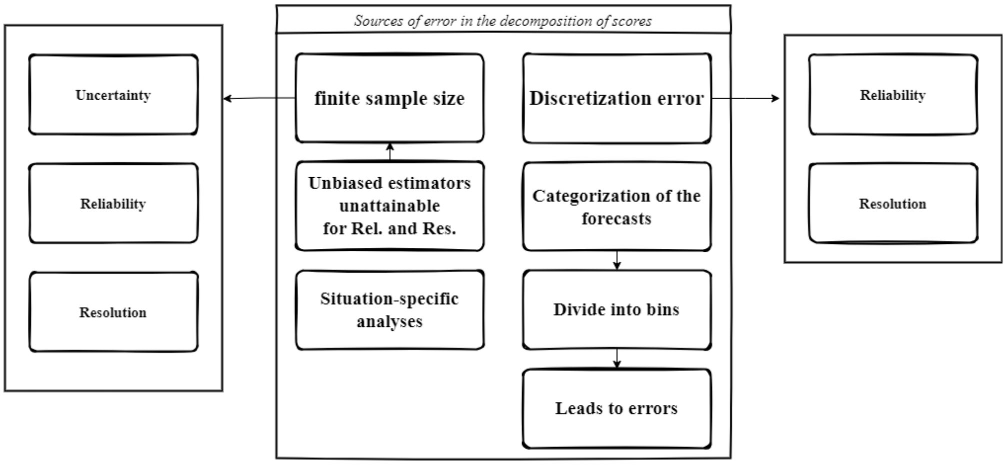

Sources of error in the decomposition of scores
Uncertainty
Reliability
Resolution
finite sample size
Unbiased estimators unattainable for Rel. and Res.
Situation-specific analyses
Discretization error
Categorization of the forecasts
Divide into bins
Leads to errors
Reliability
Resolution

Figure 3.14: The bias of the decomposition

As shown in the figure, there are two main sources of bias in score estimation: discretisation error on the one hand, and sampling uncertainty due to finite sample size on the other. The discretisation error arises from the classification of the predictions, which is necessary for the estimation of the decomposition. It affects the score as well as the reliability and resolution estimates, while the uncertainty remains constant. As in the decomposition of the quantile score presented in the previous section, it can be seen that for the analysis of uncertainty, the corresponding check-loss function focuses only on the quantile of the corresponding probability level in the validation set (observations). Therefore, the discretisation error does not affect the computation of the quantile score for uncertainty and the overall quantile score.

Other biases in the score estimates are caused by the finite sample size and affect the reliability, resolution, and uncertainty estimates, but do not affect the mean scores themselves.A detailed derivation and discussion of sample biases in the assessment of reliability and resolution estimates is provided in the appendix of the Bentzien and Friederichs 2014(Subject to discretisation with the same population bins I_k).Without further development here, this report will focus more on the impact of decomposition itself on reliabilty and resolution below.

As mentioned earlier, estimates of reliability and resolution may depend heavily on classification. To study the effect of classification, each dataset containing N forecast-observation pairs is divided into K equal intervals, each containing $N_k = N/K$ data points.The first procedure adopted was to divide the forecast probabilities into intervals of equal width and derive the observed relative frequencies of the conditions.Therefore, two methods of equidistant segmentation of the predicted values $\{p_{\tau,n}\}$ were first experimented with. prediction at a given probability level. Based on the

```
np.linspace(np.min(forecastvector), np.max(forecastvector), numbins)
```

it can divide a set of equidistant points.Where the *numbins* corresponds to the number K of bins we divide.

Equidistant centre division

One is that, according to the definition of isometric division, the boundary points or representative values of the predicted samples can be divided isometrically first. The first method uses the boundary points as the geometric distance centre of the judgement, and finds the nearest predicted values to that centre in the quantile prediction at that probability level to be assembled into a group.

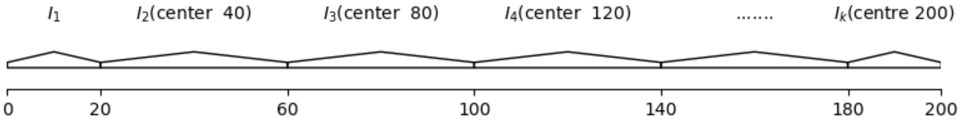


Figure 3.15: Example of the first linear division of the prediction(vec_forecast.max=200)

Equidistant boundaries division

The second is to directly divide the boundary points by equidistant division as the boundary of the grouping. Predictions that fall into the interval of two boundary points are naturally grouped together.

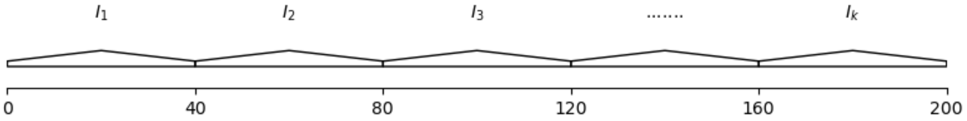


Figure 3.16: Example of the second method of division(vec_forecast.max=200)

In practice both methods work well for decomposition of QS scores in given probability levels.However, due to the low computational efficiency of the first method and the limitation of computational power (each time, one needs to judge the equidistant point on the numerical axis that is geometrically closest to one's own geometrical distance by computational traversal), thus,the integration of the QS components is based on the second method.

Quantile boundaries divisions

A review of the literature by Bentzien and Friederichs 2014 and a study by Bröcker 2008b showed that the discretisation error is reduced if the categories are chosen to contain the same number of predictor-observation pairs. They further suggest that the number of categories should be adjusted to the sample size. However, any categorisation automatically leads to errors, even for fully reliable forecasting systems.

Therefore, the definition of the subsample I_k is a key point in QS discretisation. In order to minimise the discretisation error, a new grouping method is suggested in Bentzien and Friederichs 2014.

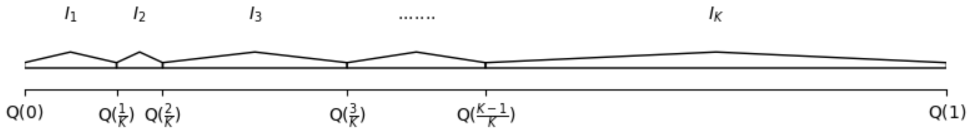


Figure 3.17: The third method of division

This approach uses a binning procedure similar to that of Atger 2004. The quantitative predictions are sorted into K equal number of intervals, i.e., the intervals are defined by the $\frac{1}{K}$ per centiles of the predictions $\{p_{\tau,n}\}$, $n = 1, ..., N$.

As shown in 3.17, an ununiform distribution of predictions can lead to under-sampling of certain categories, which can strongly bias the decomposition, so using quantiles as boundaries for dividing the bins allows for a more uniform number of predictions to be populated in each group, and has been shown to reduce the error in the previous literature on decomposition theory.

3.4.4 Experiment

According to Equation. 3.4 and 3.5, this section not only wants to achieve the categorisation decomposition of the QS score, but in addition, similar to the Brier Score approach to the study, although we do not have the exact relationship between the components of the QS score with respect to the integral against the level of probability and the components of the CRPS, we still want to obtain the range of the number of intervals to be divided with respect to the the number of corresponding CRPS(Brier) components. Therefore, we continued our previous research approach by decomposing and integrating the QS score for a given probability level, proving numercially the equivalence of Equation 3.11.

$$QS = REL_{qs} - RES_{qs} + UNC_{qs} \tag{3.16}$$

$$\int QS(x)d\tau = \int REL_{qs}(x)d\tau - \int RES_{qs}(x)d\tau + \int UNC_{qs}(x)d\tau \tag{3.17}$$

The quantile score for the ensemble forecast for a given probability level and how global information about the prediction performance of the prediction system can be obtained from the quantile score are shown in the following.

QS at different probability levels

Some users may be interested in the performance (e.g., overprediction or underprediction) of particular quantities, especially those associated with the tails of the predictive distribution. For the eight test sets with orders of magnitude in 10^4 pairs of observations, we are given $K = 1000$ intervals, where using the QS score we can get a clear picture of the prediction performance of the predictive distribution at a particular probability level, as well as the prediction performance of events at the extremes of the predictive distribution at the ends of the distribution (e.g., probability levels of 0.1 and 0.9) .

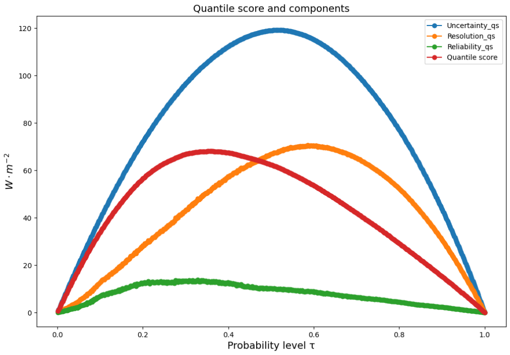


Figure 3.18: Decomposition of QS(Site MIL)

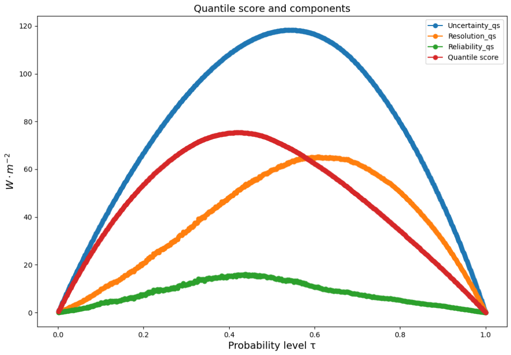
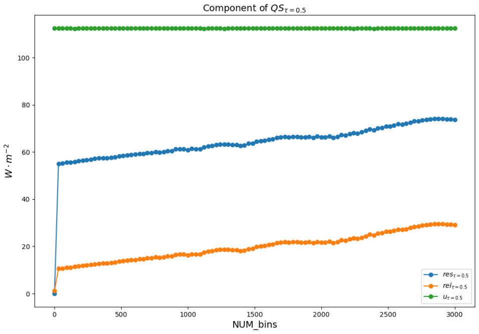


Figure 3.19: Decomposition of QS(Site PAY)

QS changes in the effect of different intervals dividing

In more detail, we next examine the effect of the QS score and its decomposition on the number of its interval divisions at a probability level of 0.5 as an example.

Discretisation does not change the overall average Quantile score, nor does it change the uncertainty (brought about by uncertainty in the distribution of sampled observations). Only reliability and resolution are affected in this process, and it can be noticed that the effect of the number of dividing intervals on the resolution will be less pronounced when a clearly meaningful (greater than 0) resolution is obtained. As resolution increases, reliability is sacrificed, so to some extent, when the number of dividing intervals is not determined, the value of reliability is not meaningful for comparison at the moement. In other words, we need to fix the same number of intervals for all sites/models when we are evaluating the prediction system using quantile scores or potential points.



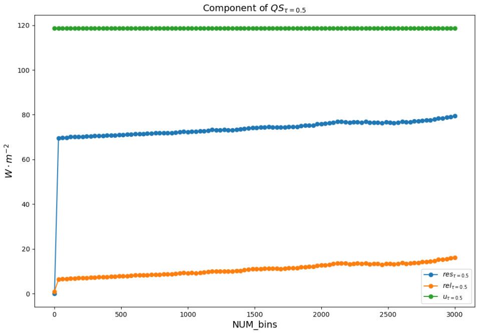


Figure 3.20: Effect of K on QS decomposition(CAB) Figure 3.21: Effect of K on QS decomposition(CAR)

As the number of dividing intervals(K) increases, the number of prediction-observation pairs corresponding to each small bins decreases, reducing the accuracy when performing the quantile estimation required for decomposition, so that if we continue to increase the number of dividing intervals, we end up with a situation in which reliability converges exactly to quantile scores and resolution converges to uncertainty, and at the other extreme, when the dividing intervals are very small, such as single-digit division intervals, in which case the difference between the quantiles we estimate from the observed validation sample and the divided quantiles is not large enough to result in 0 resolution. So it is necessary to choose an appropriate number of division intervals.

QS components and corresponding to CRPS

According to the table for integrating the QS at different probability levels, i.e., solving for the area under the curves in Figures 3.16,3.17, we use the Brier score and its component integrals with respect to the threshold variable after reaching stabilisation as a reference, where the integration step corresponding to the threshold division is fixed to be 8. We can find that, at this order of magnitude for the sample size of the eight sites, the range of results obtained when dividing the 1000 to 1500 intervals is basically the near to $CRPS_{bs}$, and more precisely, when K=1500, the component integrals of the QS with respect to the probability level can be obtained with a resolution similar to that of the $CRPS_{bs}$, and the reliability of the numerical estimates. By fixing the number of intervals, the resolution and reliability of the components of the different predictions can help us to assess and compare the quality of the predictions in the collection of predictions globally.

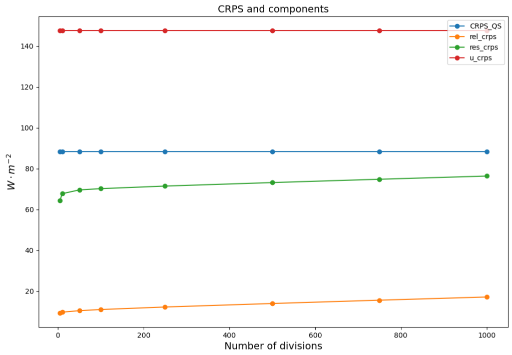


Figure 3.22: Variation of CRPS (QS) with K(PAL)

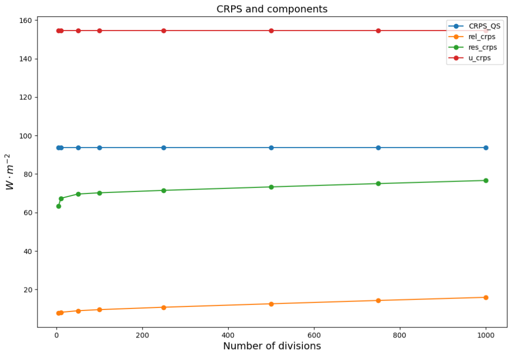


Figure 3.23: Variation of CRPS (QS) with K(PAY)

The picture 3.20 shows that since for each probability level of the prediction system the QS score exhibits the pattern we illustrated, i.e., as the division interval K increases, the resolution value gets infinitely closer to the uncertainty. This comes at the cost of an increase in the value of reliability, which consequently keeps approaching the average QS value. As for the value of CRPS it depends only on the way of dividing the probability level, since we assign the same probability weight to each sorted set member, so this value is essentially equivalent to the CRPS under the classical approach and the small integration step.

So by integrating for a uniform division of probability levels, we get the same pattern of change of CRPS components for the integrals of the corresponding components.

Site	$CRPS_{rel}(BS, intstep = 8)$	$CRPS_{rel}(QS, K = 1000)$	$CRPS_{rel}(QS, K = 1500)$
CAR	15.289890430305356	13.412010391260008	15.592876780073055
CAB	19.08594156908871	20.104812058978382	23.174019420704724
CEN	11.865166610737338	10.962780625189982	13.302738558495964
MIL	16.081641861464835	14.618006782651682	17.031040216328712
PAL	16.89856311078852	16.47479368762615	19.10470131978667
PAY	15.571788740812186	15.213112000019702	17.983001165403717
TAB	9.057628268142253	6.065609726120736	7.674859088961903
TOR	16.967228547331597	19.126853603894336	22.930085466593113

Table 3.5: Reliability integrals for Brier score and QS score

Site	$CRPS_{res}(BS, intstep = 8)$	$CRPS_{res}(QS, K = 1000)$	$CRPS_{res}(QS, K = 1500)$
CAR	93.21012780031799	91.25415776446368	93.43513375487143
CAB	77.13540045743288	78.09127142541502	81.14931718300612
CEN	75.03437559288389	74.06109719458227	76.40118023237704
MIL	86.68859216628839	85.1503128827319	87.56348126466257
PAL	76.13406620195082	75.63775873985526	78.26768684598775
PAY	76.39355198451908	75.96668161486448	78.73649043171332
TAB	90.79363509999351	87.73916165858137	89.34833730406964
TOR	69.97677169295763	72.07030418427557	75.87358948904813

Table 3.6: Resolution integrals for Brier score and QS score

Site	$CRPS(BS, intstep = 8)$	$CRPS(QS)$	$CRPS_{unc}(BS, intstep = 8)$	$CRPS_{unc}(QS)$
CAR	78.87701790208784	78.90760311696165	156.79725527210053	156.74336151101036
CAB	89.04396213227595	89.06712425877623	147.09342102062016	147.04193840235135
CEN	86.01066995304916	86.04581526620015	149.17987893519577	149.13329261900336
MIL	86.19049279404976	86.24382983077493	156.79744309887337	156.76493786248153
PAL	88.35349595407834	88.40011860797651	147.58899904524074	147.54988845700998
PAY	93.79051699939455	93.83606107951363	154.61228024310134	154.5756311526476
TAB	62.74018039309199	62.73434646493634	144.47618722494332	144.40215295344225
TOR	78.87444210989734	78.87980887568605	131.8839852555232	131.79373991833654

Table 3.7: Intergrals for uncertainties and CRPS values

In sum, by integrating the components of QS, we can get CRPS and also get almost accurate estimate of CRPS uncertainty, but for reliability and resolution, for different datasets we need to test to select appropriate K values for comparison and judgment.

3.5 Parametric distributions

During the internship, since python's computational library for probability measures only covers assumptions about parameter distributions based on the Gaussian distribution, even though it is widely used. However, in many application scenarios other distributions need to be applied, such as predicting some extreme distributions for weather(Gumbel,Weibull,etc), and gamma distribution for predicting precipitation, etc., so the existing library lacks these functions.

Therefore, through reading the literature, understanding the theoretical derivation of analytic expressions by verifying the quantile function of the distribution of interest,as well as collecting the corresponding CRPS expressions under the prediction assumptions of the relevant parameter distributions, we have completed the collation and writed python functions of CRPS for some relevant parameter distributions. Because this part of the work against the R language has been quite complete, so this part will not be expanded.(Refer to Appendix)

Conclusion

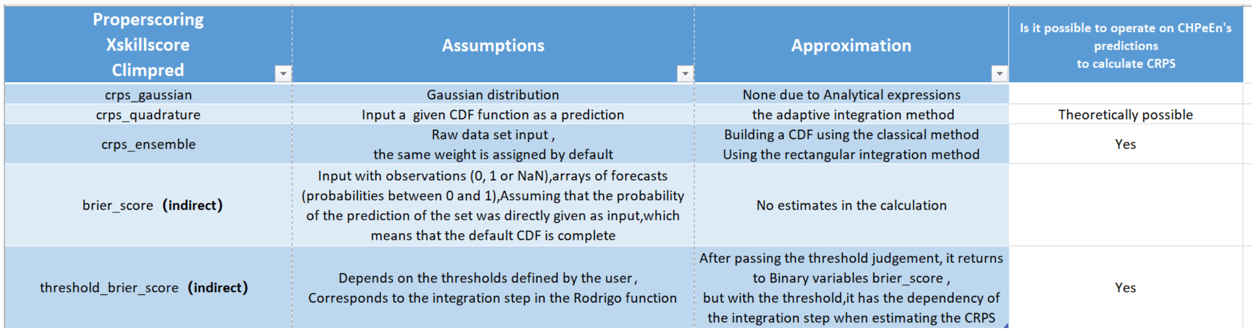
During my twenty-week internship, I was able to develop a theoretical understanding of the field of solar energy forecasting from 0 to 1 and gained an awareness of quality assessment methods for solar energy forecasting through relevant mathematical knowledge. By choosing this field of my interest, it laid a very good foundation for my research study in China afterwards and also give me the possibility to think about where I want to explore from here. In addition, the internship was oriented to a relevant research project, which also allowed me to understand the implementation of multiple processes of an practical/academic topic from start to finish, defining the issues to be addressed, formulation, feasibility analysis, etc.

In terms of specific work, the refinement of the Brier score decomposition method for calculating CRPS components was completed, specifying the integration methods that should be taken under different CDFs. Completed the practice and effect analysis of the decomposition of QS score, i.e., a suitable number of interval divisions needs to be selected to obtain relatively meaningful reliability and resolution indexes. And the equivalence of the uncertainties obtained by the two methods further proves the theory of the decomposition of the proper score.

In addition, by participating in a part of the project, part of the work may be used later in a teaching tool for the jupyter book as well as participating in the writing of data related type articles. Due to time constraints and my productivity issues, I did not do a perfect job as expected, accomplishing all of the intended tasks. However, it was a very good start, and the twenty weeks of internship in Réunion gave me a better practical knowledge of the field of solar energy forecasting, and I am sure that I will continue to work in a relevant context afterwards.

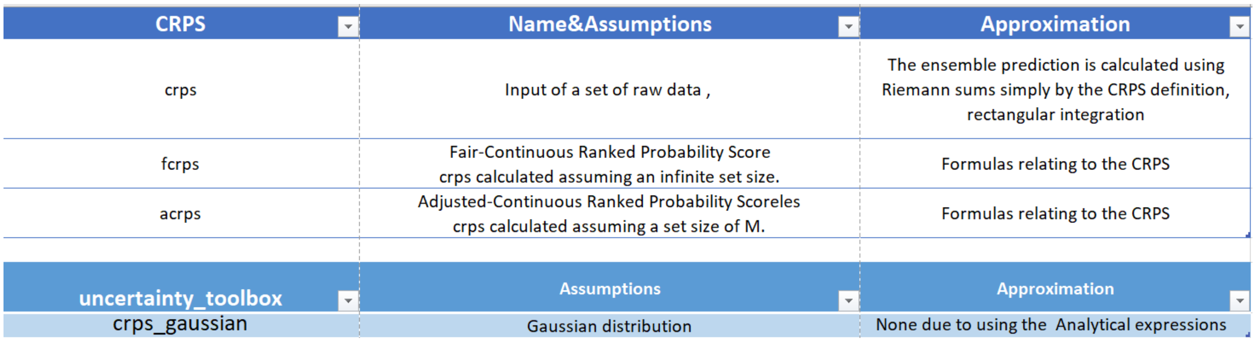
Appendix

3.6 Summary of python's existing CRPS computing libraries



Properscoring Xskillscore Climpred	Assumptions	Approximation	Is it possible to operate on CHPeEn's predictions to calculate CRPS
crps_gaussian	Gaussian distribution	None due to Analytical expressions	
crps_quadrature	Input a given CDF function as a prediction	the adaptive integration method	Theoretically possible
crps_ensemble	Raw data set input , the same weight is assigned by default	Building a CDF using the classical method Using the rectangular integration method	Yes
brier_score **(indirect)**	Input with observations (0, 1 or NaN),arrays of forecasts (probabilities between 0 and 1),Assuming that the probability of the prediction of the set was directly given as input,which means that the default CDF is complete	No estimates in the calculation	
threshold_brier_score **(indirect)**	Depends on the thresholds defined by the user , Corresponds to the integration step in the Rodrigo function	After passing the threshold judgement, it returns to Binary variables brier_score , but with the threshold,it has the dependency of the integration step when estimating the CRPS	Yes

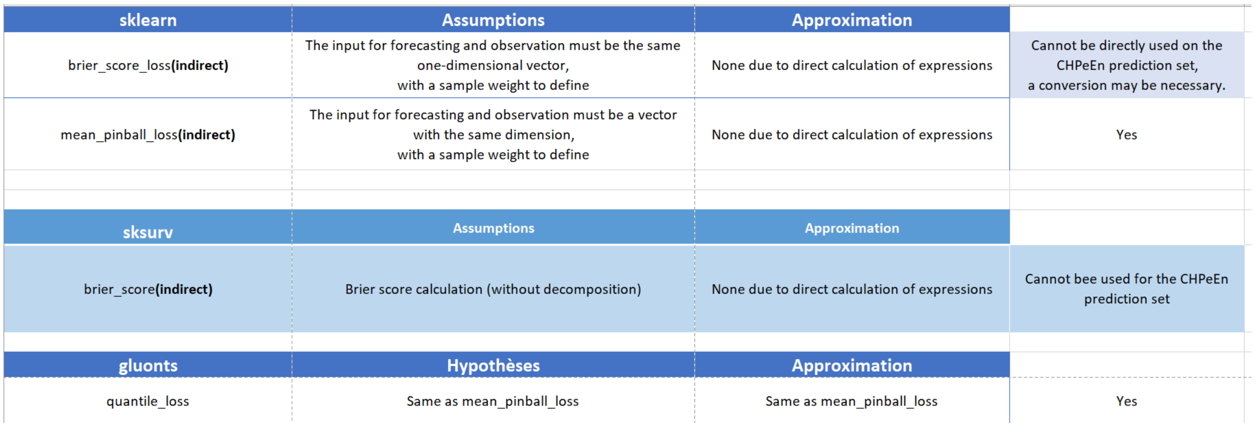
Figure 3.24: The bias of the decomposition



CRPS	Name&Assumptions	Approximation
crps	Input of a set of raw data ,	The ensemble prediction is calculated using Riemann sums simply by the CRPS definition, rectangular integration
fcrps	Fair-Continuous Ranked Probability Score crps calculated assuming an infinite set size.	Formulas relating to the CRPS
acrps	Adjusted-Continuous Ranked Probability Scoreles crps calculated assuming a set size of M.	Formulas relating to the CRPS

uncertainty_toolbox	Assumptions	Approximation
crps_gaussian	Gaussian distribution	None due to using the Analytical expressions

Figure 3.25: The bias of the decomposition



sklearn	Assumptions	Approximation	
brier_score_loss**(indirect)**	The input for forecasting and observation must be the same one-dimensional vector, with a sample weight to define	None due to direct calculation of expressions	Cannot be directly used on the CHPeEn prediction set, a conversion may be necessary.
mean_pinball_loss**(indirect)**	The input for forecasting and observation must be a vector with the same dimension, with a sample weight to define	None due to direct calculation of expressions	Yes
sksurv	**Assumptions**	**Approximation**	
brier_score**(indirect)**	Brier score calculation (without decomposition)	None due to direct calculation of expressions	Cannot bee used for the CHPeEn prediction set
gluonts	**Hypothèses**	**Approximation**	
quantile_loss	Same as mean_pinball_loss	Same as mean_pinball_loss	Yes

Figure 3.26: The bias of the decomposition

3.7 Equivalence with classical CRPS and CRPS via QS integrals

Reminder of the problem

$$\begin{cases} CRPS(\hat{F},x) = \int_{-\infty}^{\infty}[\hat{F}(y) - H(y-x)]^2 dy \\ CRPS(\hat{F},x) = 2\int_0^1 QS_\tau d\tau \end{cases}$$

Where QS$_\tau$ $= \tau * H(F_\tau < x) * (x - F_\tau) + (1-\tau) * H(F_\tau > x) * (F_\tau - x)$

Proof

Proper score character

Multiply by 2 and rewrite the definition of Quantile Score as follows,

$$2QS_\tau = |x - F_\tau| + (2\tau - 1) * (x - F_\tau) \tag{3.18}$$

By defining the QS (Quantile Score) as a proper score, it is known that the minimum is considered to be exactly the same as the observed and predicted distributions, and we therefore want to minimise this expectation. Let x be the set of all possible future predicted values to be taken, or $f(x)$ is the probability density function(PDF) of the actual future distribution, i.e. the sample distribution for validation.

$$E(2QS_\tau) = \int_{-\infty}^{\infty} 2QS_\tau f(x)dx \tag{3.19}$$

Equation (1) is substituted for equation (2), which is expressed as follows,

$$(2)=\int_{-\infty}^{\infty}\left(|x - F_\tau| + (2\tau - 1) * (x - F_\tau)\right) f(x)dx$$

i.e.,
$E(2QS_\tau) = \int_{-\infty}^{\infty} 2QS_\tau f(x)dx = 2\tau \int_{-\infty}^{\infty} (x - F_\tau) f(x)dx + 2\int_{-\infty}^{F_\tau}(F_\tau - x)f(x)dx$

$$= -2\tau(F_\tau - \int_{-\infty}^{\infty} xf(x)dx) + 2\int_{-\infty}^{F_\tau}(F_\tau - x)f(x)dx$$

Derived from this expectation,
$\frac{dE}{dF_\tau} = -2\tau + [2\int_{-\infty}^{F_\tau} F_\tau f(x)dx]' - 2F_\tau f(F_\tau)$
$= -2\tau - 2F_\tau f(F_\tau) + 2F_\tau f(F_\tau) + 2\int_{-\infty}^{F_\tau} f(x)dx$ Take the extreme point where the derivative is equal to 0, $\tau = \int_{-\infty}^{F_\tau} f(x)dx$,

According to the definition of the probability density function,The CDF of the predictive distribution is defined here as $\hat{F}(x)$, The PDF corresponding to the predicted distribution is expressed as$\hat{f}(x)$,When the observed distribution is identical to the predicted distribution, i.e. when $\hat{f}(x) = f(x)$, we have,
$\tau = \int_{-\infty}^{F_\tau} f(x)dx = \hat{F}(F_\tau)$
$F_\tau = \hat{F}^{-1}(\tau)$

Thus, when the forecast is made at this point, the expectation of the QS score has a locally minimal/optimal value. So at this point, the definition of the eigen score is always respected, Let's now return to the definition of the QS score to calculate the CRPS.

Proof of CRPS equivalence

According to equation (1) in part 2.1

We have shown that for a QS, it is only a clean score if the prediction is satisfied $F_{tau} = hatF^{-1}(\tau)$, so we have,

$$\int_0^1 2QS_\tau d\tau = \int_0^1 |x - \hat{F}^{-1}(\tau)| + (2\tau - 1) * \big(x - \hat{F}^{-1}(\tau)\big)d\tau \tag{3.20}$$

Here we carry out an integral variable substitution which $y = \hat{F}^{-1}(\tau)$,c'est-à-dire $\tau = \hat{F}(y)$

so we have ,$d\tau = d\hat{F}(y) = \hat{f}(y)dy$ here, $\hat{F}(y)$ et $\hat{f}(y)$ represent respectively the CDF and the PDF of the predicted distribution.

by replacing the upper and lower limits of integration and the variable of integration, the result is,

$$\int_0^1 2QS_\tau d\tau = \int_{-\infty}^{\infty} \big[|y - x| + \big(1 - 2\hat{F}(y)\big)(y - x)\big]\hat{f}(y)dy \tag{3.21}$$

where (4) can be expressed as follows,

$$\int_{-\infty}^{\infty} 2(H\big(y - x) - \hat{F}(y)\big)(y - x)\hat{f}(y)dy$$

The integration by parts operations are performed as follows,
$\int_{-\infty}^{\infty} 2(H\big(y - x) - \hat{F}(y)\big)(y - x)\hat{f}(y)dy = \int_{-\infty}^{\infty} 2(H\big(y - x) - \hat{F}(y)\big)(y - x)d\hat{F}(y) =$

$$\underbrace{\big[2\big(H(y - x) - \hat{F}(y)\big)(y - x)\hat{F}(y)\big]_{-\infty}^{\infty}}_{\text{⓪}} \underbrace{- 2\int_{-\infty}^{\infty} \hat{F}(y)d(H\big(y - x) - \hat{F}(y)\big)(y - x)}_{\text{①}}$$

The differential $textcircled1$ terms in the second part below can be broken down as follows,
where

$$\underbrace{- 2\int_{-\infty}^{\infty} \hat{F}(y)d(H\big(y - x) - \hat{F}(y)\big)(y - x)}_{\text{①}}$$

$$=\underbrace{2\int_{-\infty}^{\infty}\hat{F}(y)(y-x)(\hat{f}(y)-\delta(y-x))dy}_{\textcircled{2}}\underbrace{-2\int_{-\infty}^{\infty}\hat{F}(y)\big(H(y-x)-\hat{F}(y)\big)dy}_{\textcircled{3}}$$

where the first term of the equation can be expressed as follows:

$$\underbrace{2\int_{-\infty}^{\infty}\hat{F}(y)(y-x)(\hat{f}(y)-\delta(y-x))dy}_{\textcircled{2}}=\underbrace{-2\int_{-\infty}^{\infty}\hat{F}(y)(y-x)\delta(y-x)dy}_{\textcircled{4}}+\underbrace{2\int_{-\infty}^{\infty}\hat{F}(y)(y-x)\hat{f}(y)dy}_{\textcircled{5}}$$

The original integral can therefore be expressed as follows, $\int_0^1 2QS\tau \mathrm{d}\tau = \textcircled{0}+\textcircled{3}+\textcircled{4}+\textcircled{5}$

⓪

$\big[2\big(H(y-x)-\hat{F}(y)\big)(y-x)\hat{F}(y)\big]_{-\infty}^{\infty}$

Consider that when $y \to \pm\infty$, on a, $H(y-x)-\hat{F}(y)=0$,

④

-2$\int_{-\infty}^{\infty}\hat{F}(y)(y-x)]\delta(y-x)dy$

Consider that when $\int_{-\infty}^{\infty} f(t)\delta(t-T)dt = f(T)$,

so we have, $\int_{-\infty}^{\infty}\hat{F}(y)(y-x)\delta(y-x)dy = \hat{F}(y)(y-y)=0$

③

-2$\int_{-\infty}^{\infty}\hat{F}(y)\big(H(y-x)-\hat{F}(y)\big)dy$

⑤

$+2\int_{-\infty}^{\infty}\hat{F}(y)(y-x)\hat{f}(y)dy$

for ⑤,Using the law of partial integration,

$$2\int_{-\infty}^{\infty}\hat{F}(y)(y-x)\hat{f}(y)dy = 2\int_{-\infty}^{\infty}\hat{F}(y)(y-x)d\hat{F}(y)$$
$$= 2\hat{F}^2(y)(y-x)_{-\infty}^{\infty} - 2\int_{-\infty}^{\infty}\hat{F}(y)\big[(y-x)\hat{f}(y)+\hat{F}(y)\big]dy$$

Thus,$2\int_{\infty}^{\infty}\hat{F}(y)(y-x)\hat{f}(y)dy = \hat{F}^2(y)(y-x)_{-\infty}^{\infty} - \int_{-\infty}^{\infty}\hat{F}^2(y)dy$

Thus,③ + ⑤,on a $-\int_{-\infty}^{\infty}\hat{F}^2(y)dy - 2\int_{-\infty}^{\infty}\hat{F}(y)\big(H(y-x)-\hat{F}(y)\big)dy + \int_{-\infty}^{\infty}H(y-x)dy$

$=\int_{-\infty}^{\infty}[\hat{F}(y)-H(y-x)]^2dy$

All it takes is to prove that,$\int_{-\infty}^{\infty}H(y-x)dy = \hat{F}^2(y)(y-x)_{-\infty}^{\infty}$ This theoretically demonstrates the equivalence of the CRPS definition calculation method and the QS score calculation method.

To summarise, according to the above decomposition,

equation (4) =⓪ + ① = ⓪ + ② + ③ = ⓪ + ③ + ④ + ⑤

where ⓪ et ④ are equal to 0 . The following equivalence is obtained using the result of ③ + ⑤,

$$\int_0^1 2QS_\tau d\tau = \int_{-\infty}^{\infty}\big[|y-x|+\big(1-2\hat{F}(y)\big)(y-x)\big]\hat{f}(y)dy = \int_{\infty}^{\infty}[\hat{F}(y)-H(y-x)]^2dy \tag{3.22}$$

Bibliography

Le Gal La Salle, Josselin (July 2021). "Qualité et valeur des prévisions solaires probabilistes". PhD thesis.

Yang, Dazhi et al. (May 2020). "Verification of deterministic solar forecasts". In: *Solar Energy* 210. DOI: 10.1016/j.solener.2020.04.019.

Lauret, Philippe, Mathieu David, and Pierre Pinson (2019). "Verification of solar irradiance probabilistic forecasts". In: *Solar Energy* 194, pp. 254–271. ISSN: 0038-092X. DOI: https://doi.org/10.1016/j.solener.2019.10.041. URL: https://www.sciencedirect.com/science/article/pii/S0038092X19310382.

Hamill, Thomas M. (2001). "Interpretation of Rank Histograms for Verifying Ensemble Forecasts". In: *Monthly Weather Review* 129, pp. 550–560.

Pinson, Pierre et al. (2007). "Non-parametric probabilistic forecasts of wind power: required properties and evaluation". In: *Wind Energy* 10, pp. 497–516.

Hersbach, Hans (Oct. 2000). "Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems". In: *Weather and Forecasting - WEATHER FORECAST* 15, pp. 559–570. DOI: 10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2.

Brier, Glenn W. (1950). "VERIFICATION OF FORECASTS EXPRESSED IN TERMS OF PROBABILITY". In: *Monthly Weather Review* 78, pp. 1–3. URL: https://api.semanticscholar.org/CorpusID:122906757.

Brown, Thomas A. (1970). *Probabilistic Forecasts and Reproducing Scoring Systems*. Santa Monica, CA: RAND Corporation.

Good, I. J. (1952). "Rational Decisions". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 14.1, pp. 107–114. DOI: https://doi.org/10.1111/j.2517-6161.1952.tb00104.x. eprint: https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/j.2517-6161.1952.tb00104.x. URL: https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1952.tb00104.x.

Savage, Leonard J. (1971). "Elicitation of Personal Probabilities and Expectations". In: *Journal of the American Statistical Association* 66.336, pp. 783–801. ISSN: 01621459. URL: http://www.jstor.org/stable/2284229 (visited on 08/07/2023).

Schervish, Mark J. (1989). "A General Method for Comparing Probability Assessors". In: *The Annals of Statistics* 17.4, pp. 1856–1879. DOI: 10.1214/aos/1176347398. URL: https://doi.org/10.1214/aos/1176347398.

Gneiting, Tilmann and Adrian E Raftery (2007). "Strictly Proper Scoring Rules, Prediction, and Estimation". In: *Journal of the American Statistical Association* 102.477, pp. 359–378. DOI: 10.1198/016214506000001437. eprint: https://doi.org/10.1198/016214506000001437. URL: https://doi.org/10.1198/016214506000001437.

Bentzien, Sabrina and Petra Friederichs (2014). "Decomposition and graphical portrayal of the quantile score". In: *Quarterly Journal of the Royal Meteorological Society* 140.683, pp. 1924–1934. DOI: https://doi.org/10.1002/qj.2284. eprint: https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/qj.2284. URL: https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.2284.

Yang, Dazhi (2019). "A universal benchmarking method for probabilistic solar irradiance forecasting". In: *Solar Energy* 184, pp. 410–416. ISSN: 0038-092X. DOI: https://doi.org/10.1016/j.solener.2019.04.018. URL: https://www.sciencedirect.com/science/article/pii/S0038092X19303457.

Bröcker, Jochen (2008a). "Decompositions of Proper Scores". In: URL: https://api.semanticscholar.org/CorpusID:17259047.

Laio, Francesco and Stefania Tamea (Aug. 2006). "Verification tools for probabilistic forecasts of continuous hydrological variables". In: *Hydrology and Earth System Sciences* 11. DOI: 10.5194/hessd-3-2145-2006.

Bröcker, Jochen (2008b). “Some Remarks on the Reliability of Categorical Probability Forecasts”. In: *Monthly Weather Review* 136, pp. 4488–4502. URL: https://api.semanticscholar.org/CorpusID:122132518.

Atger, Frédéric (2004). “Estimation of the reliability of ensemble-based probabilistic forecasts”. In: *Quarterly Journal of the Royal Meteorological Society* 130.597, pp. 627–646. DOI: https://doi.org/10.1256/qj.03.23. eprint: https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1256/qj.03.23. URL: https://rmets.onlinelibrary.wiley.com/doi/abs/10.1256/qj.03.23.