Australian Government
Bureau of Meteorology

Estimation of CRPS for precipitation forecasts using weighted sums of quantile scores and Brier scores

Robert Taggart

March 2023

Bureau Research Report - 079

Estimation of CRPS using weighted sums of quantile scores and Brier scores

Robert Taggart

Bureau Research Report No. 079

March 2023

National Library of Australia Cataloguing-in-Publication entry

Authors: Robert Taggart

Title: Estimation of CRPS for precipitation forecasts using weighted sums of quantile scores and Brier scores

ISBN: 978-1-925738-67-4

ISSN: 2206-3366

Series: Bureau Research Report – BRR079

Enquiries should be addressed to:

Lead Author: Robert Taggart

Bureau of Meteorology

PO Box 413,

Darlinghurst, NSW 1300, Australia

robert.taggart@bom.gov.au:

Copyright and Disclaimer

Contents

List of Figures

List of Tables

Executive Summary

The continuous ranked probability score (CRPS) is a commonly used strictly proper scoring rule to measure the forecast accuracy of predictive distributions. Since the CRPS can be expressed as an integral of Brier scores, it can be approximated using a weighted sum of a finite number of Brier scores. Similarly, the CRPS can be expressed as an integral of quantile scores and hence can be approximated using a weighted sum of a finite number of quantile scores. This research report examines how well the CRPS can be approximated for operational predictive distributions of precipitation using (a) quantile scores at four specific quantile levels, (b) Brier scores associated with seven specific event thresholds and (c) combinations of the above. The specific quantile levels and event thresholds are those that are published by the Australian Bureau of Meteorology. It is found that weighted sums of the four quantile scores estimate the CRPS with a very high degree of accuracy. Such weighted sums are themselves strictly proper scoring rules and can be used as a substitute to the CRPS when limited information about the predictive distribution is available, or when ease of calculation is desired.

1. Introduction

The continuous ranked probability score (CRPS) is a commonly used scoring rule for assessing the predictive performance of predictive distributions (Matheson and Winkler 1976). Suppose that a predictive distribution F has the form of a cumulative distribution function (CDF), so that for the unknown quantity Y being forecast,

$$\mathbf{P}(Y \leq z) = F(z)$$

for any real number z. Then the CRPS of F for corresponding real valued observation y is given by

$$\mathrm{CRPS}(F, y) = \int_{-\infty}^{\infty} (F(\theta) - \mathbf{1}\{y \leq \theta\})^2 \, \mathrm{d}\theta, \qquad (1)$$

where the indicator function $\mathbf{1}$ has the property that $\mathbf{1}\{y \leq \theta\}$ takes the value 1 whenever $y \leq \theta$ and is 0 otherwise. The units of the CRPS are the same as the observation. CRPS values are non-negative and a lower CRPS indicates a better forecast. By converting a single-valued forecast into a CDF that has the form of an indicator function, it is seen that the CRPS is a generalisation of the absolute error for single-valued forecasts. The CRPS is a strictly proper scoring rule, meaning that a forecaster will minimise their expected CRPS only by forecasting the distribution that aligns with their best judgment. Apart from strict propriety, another reason for assessing predictive performance using the CRPS is that it weights predictive performance for all user decision thresholds θ equally (Gneiting and Ranjan 2011). This contrasts with the ranked probability score (RPS), which only considers predictive performance for each decision threshold associated with each term in a finite sum.

In practice, the CRPS may be difficult or expensive to calculate. If only a small number of values of the predictive CDF are known then it may be difficult to reconstruct the CDF with sufficient accuracy, casting doubt on the calculated value of the CRPS. If many values of the CDF are known or if there are many forecast cases then CRPS calculations can be computationally expensive in memory or time, particularly for applications where bulk statistics are calculated on the fly. Interactive verification dashboards developed by the Australian Bureau of Meteorology encounter such issues when the scoring predictive performance of precipitation forecasts published by the Bureau to the Australian Digital Forecast Database (ADFD). On the one hand, no more than 11 points on the predictive CDFs are known for daily precipitation forecasts, and no more than 4 points are known for 3-hourly precipitation forecasts. On the other hand, millions of forecast cases are published each day.

Fortunately, these known points from the Bureau's predictive CDFs can be expressed either as probability of non-exceedance forecasts for a fixed set of non-exceedance thresholds, or as quantile forecasts for a fixed set of quantile levels. Such forecasts can be efficiently scored using the Brier score and quantile score respectively. This report shows how weighted sums of such scores give reasonable to excellent estimates for the CRPS. Moreover, these weighted sums of scores are themselves strictly proper scoring rules, so that they cannot be hedged.

The report is structured as follows. Section 2 gives a theoretical explanation why the CRPS can be estimated using weighted sums of Brier scores and quantile scores, for some suitable choice of weights. The weights will be selected empirically using a dataset of operational forecasts as follows. First, the Brier scores, quantile scores and CRPS will be calculated for each forecast case in the dataset. To calculate the CRPS, each CDF will be reconstructed using known points from the predictive distribution via a method described in Section 3. Second, these CRPS values are predicted using linear models with Brier scores and quantile scores as predictors, as described in Section 4. The accuracy of these models, their coefficients and sensitivity to overfitting will be discussed in Section 5. Finally, Section 6 summarises the results and recommends two proper scoring rules, constructed using weighted sums of quantile scores, as suitable substitutes for the CRPS for the applications considered in this research report.

2. The CRPS as an integral of Brier and quantile scores

The Brier score is a strictly proper scoring rule for assessing probability of non-exceedance forecasts. Given a real valued threshold θ, an event is said to have occurred if the real valued observation y satisfies $y \leq \theta$, and to have not occurred if $y > \theta$. Given a predictive CDF F, the forecast probability of an event occurring is $F(\theta)$. The Brier score (BS) for the forecast $F(\theta)$ and corresponding binary observation $\mathbf{1}\{y \leq \theta\}$ is given by

$$\mathrm{BS}(F(\theta), \mathbf{1}\{y \leq \theta\}) = (F(\theta) - \mathbf{1}\{y \leq \theta\})^2. \qquad (2)$$

It is apparent from Equations (1) and (2) that the CRPS is an integral of Brier scores over all event thresholds θ, which suggests that

$$\mathrm{CRPS}(F, y) \approx \sum_{i=1}^{n} c_i \, \mathrm{BS}(F(\theta_i), \mathbf{1}\{y \leq \theta_i\}) \qquad (3)$$

for an increasing sequence of real numbers θ_i and suitable non-negative constants c_i.

The quantile score is a strictly proper scoring rule for quantile forecasts. Given a real valued forecast x and corresponding real valued observation y, the quantile score $\mathrm{QS}_\alpha(x, y)$ at level α, where $0 < \alpha < 1$, is defined by

$$\mathrm{QS}_\alpha(x, y) = \begin{cases} \alpha|x - y| & \text{if } x < y \\ (1 - \alpha)|x - y| & \text{if } x \geq y. \end{cases}$$

It is called the quantile score because a single-valued forecast x that optimises the forecaster's expected score is an α-quantile of the forecaster's predictive distribution F (Gneiting and Raftery 2007). Laio and Tamea (2007) showed that the CRPS can be expressed as an integral of quantile scores:

$$\mathrm{CRPS}(F, y) = 2 \int_0^1 \mathrm{QS}_\alpha(\mathrm{Q}_\alpha(F), y) \, \mathrm{d}\alpha,$$

where $\mathrm{Q}_\alpha(F)$ is an α-quantile of F. Thus, for suitable non-negative constants k_i associated with quantile levels α_i,

$$\mathrm{CRPS}(F, y) \approx \sum_{i=1}^{n} k_i \, \mathrm{QS}_{\alpha_i}\big(\mathrm{Q}_{\alpha_i}(F), y\big). \qquad (4)$$

Note that the right-hand sides of Equations (3) and (4) are weighted sums of proper scores and hence are also proper.

The remainder of the paper is devoted to finding suitable values for the constants c_i and k_i given thresholds θ_i and levels α_i that are relevant to precipitation forecasts issued by the Bureau. While these could be estimated based on the spacing of the thresholds θ_i and levels α_i, we instead use an empirical approach using reconstructed predictive distributions.

3. Reconstructing CDFs using points from the predictive distribution

For each daily validity period and grid cell, the ADFD publishes eleven precipitation forecasts that can be used to reconstruct the predictive CDF. These are listed in Table 1. For convenience, the first seven forecasts in this table will be collectively referred to as DailyPoPX forecasts, and the remaining four forecasts as DailyPrecipYPct forecasts. Together, these forecasts specify up to eleven unique points on the graph of the predictive CDF, though there may be fewer than eleven if, for example, any quantile forecast is 0 mm. The right-most column of Table 1 gives the specific values of predictive CDF issued for Sydney Airport for 8 February 2020 with a lead time of four days.

Using this information, one can reconstruct the values of the CDF between these points using linear interpolation. In most situations, the probability of exceeding 50 mm is 0%, and extrapolation of the tail of the CDF is not required. However, when DailyPrecip10Pct is greater than 50 mm, the 0.95, 0.98 and 0.99 quantiles of the CDF are extrapolated from the 0.75 and 0.90 quantiles by modelling the tail with a Weibull distribution.[1] Finally, the tail of the predictive CDF is completed using linear interpolation between these percentiles and linear extrapolation beyond the 99th percentile, clipped to the value 1.

Figure 1 illustrates the reconstruction of a predictive CDF for the official Bureau forecast for Sydney Airport on 8 February 2020, with a lead time of four days. In this case,

[1] That is, the shape and scale parameters of the Weibull distribution are found using the 0.75- and 0.90-quantiles and solving simultaneous equations. The 0.95-, 0.98- and 0.99-quantiles are then calculated using this distribution. The Weibull distribution is used here, rather than say a gamma distribution, because the Bureau's forecast production process sometimes uses a Weibull distribution, particularly for heavy rainfall situations.

DailyPoP50 = 25% and DailyPrecip10Pct = 89 mm. The DailyPoP50 and DailyPrecip25Pct forecasts give the same point on the graph.

Table 1: Daily precipitation forecasts published to the ADFD. The right-most column gives values for the Official lead day 4 forecast at Sydney Airport for 8 February 2020. This CDF is graphed in Figure 1.

Forecast name	**Description**	**Relationship to CDF F**	**Values for the specific CDF F of Figure 1**
DailyPoP	Chance (%) of any precipitation	$F(0) = 1 - \text{DailyPoP}/100$	$F(0) = 0.096$
DailyPoP1	Chance (%) of precipitation exceeding 1 mm	$F(1) = 1 - \text{DailyPoP1}/100$	$F(1) = 0.104$
DailyPoP5	Chance (%) of precipitation exceeding 5 mm	$F(5) = 1 - \text{DailyPoP5}/100$	$F(5) = 0.13$
DailyPoP10	Chance (%) of precipitation exceeding 10 mm	$F(10) = 1 - \text{DailyPoP10}/100$	$F(10) = 0.29$
DailyPoP15	Chance (%) of precipitation exceeding 15 mm	$F(15) = 1 - \text{DailyPoP15}/100$	$F(15) = 0.42$
DailyPoP25	Chance (%) of precipitation exceeding 25 mm	$F(25) = 1 - \text{DailyPoP25}/100$	$F(25) = 0.56$
DailyPoP50	Chance (%) of precipitation exceeding 50 mm	$F(50) = 1 - \text{DailyPoP50}/100$	$F(50) = 0.75$
DailyPrecip75Pct	0.25-quantile forecast	$F(\text{DailyPrecip75Pct}) = 0.25$ provided $\text{DailyPrecip75Pct} > 0$	$F(9.2) = 0.25$
DailyPrecip50Pct	0.5-quantile forecast	$F(\text{DailyPrecip50Pct}) = 0.5$ provided $\text{DailyPrecip50Pct} > 0$	$F(20.4) = 0.5$
DailyPrecip25Pct	0.75-quantile forecast	$F(\text{DailyPrecip25Pct}) = 0.75$ provided $\text{DailyPrecip25Pct} > 0$	$F(50) = 0.75$
DailyPrecip10Pct	0.9-quantile forecast	$F(\text{DailyPrecip10Pct}) = 0.9$ provided $\text{DailyPrecip10Pct} > 0$	$F(89) = 0.9$

To efficiently calculate the CRPS over many forecast cases, representation of predictive CDFs is vectorized. To manage memory issues, prior to reconstructing CDFs and calculating CRPSs, observations and quantile forecasts are rounded to the nearest 0.2 mm up to 15mm, then to the nearest 1 mm up to 100 mm, then to the nearest 5 mm up to 200 mm, then to the nearest 10 mm up to 1000 mm, then to the nearest 25 mm up to 1500 mm. Tests show that the main results of this research report are not sensitive to such rounding.

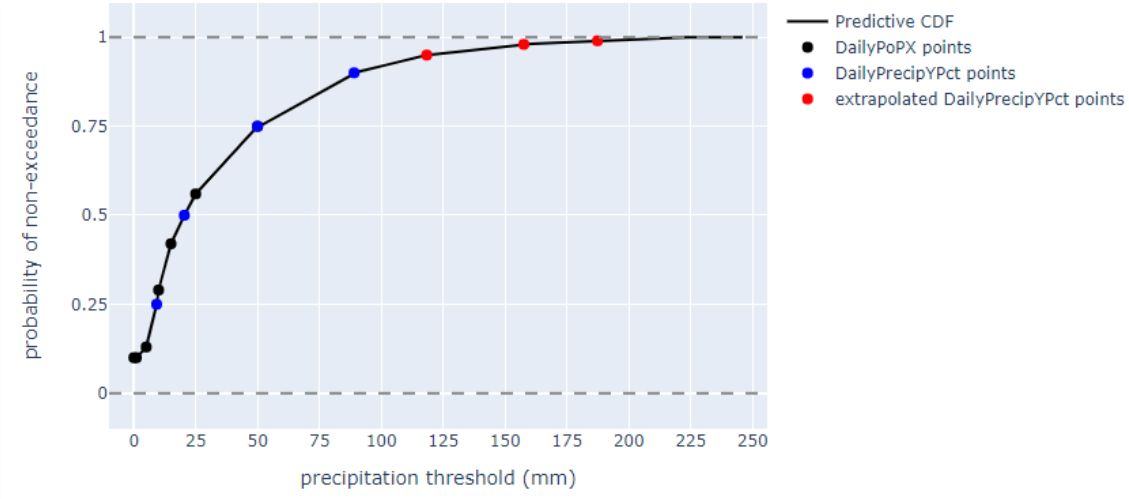

Figure 1: Reconstruction of a predictive CDF from DailyPoPX and DailyPrecipYPct forecasts.

4. Estimating weights using least squares regression

The CRPS, and corresponding quantile and Brier scores associated with DailyPoPX and DailyPrecipYPct forecasts, were calculated for daily precipitation forecasts issued by two forecast systems for 535 locations across Australia, for 730 daily validity periods (1 January 2021 to 31 December 2022) with lead times of 1 to 7 days. The two forecast systems are the Official Bureau forecasts with an afternoon issue time, and the Bureau's automated forecast system, known as AutoFcst, with a base time of 12UTC. Official forecasts are sometimes identical to AutoFcst but will differ when operational meteorologists choose to deviate from AutoFcst, particularly at shorter lead times. Excluding instances where the CRPS could not be calculated due to missing data (forecast or observed), this represents 5,254,275 forecast cases. Brier scores were also calculated for each of the corresponding probability of non-exceedance forecasts derived from DailyPoPX forecasts, and quantile scores were calculated for each of the DailyPrecipYPct forecasts.

Given a training dataset of predictors (such as the Brier scores for non-exceedance forecasts at thresholds 0 mm, 1 mm, 5 mm, 10 mm, 15 mm, 25 mm and 50 mm), the predictand (CRPS) was predicted using a linear model with coefficients determined via least squares fit; specifically, the CRPS is predicted using a linear combination of the predictors with non-negative coefficients and a constant term of 0. The predictive performance of the model was assessed using the coefficient of determination R^2, which can be interpreted as the fraction of variability in CRPS values that can be explained by the model. A higher value of R^2 indicates a better model and the highest possible value is 1. An R^2 value of 0 indicates that the model has predictive power on par with using the mean CRPS of the training dataset as the prediction of CRPS for each forecast case.

To test for the possibility of model overfitting, models were rerun using a smaller set of training data and then validated using a non-overlapping set of validation data.

Table 2 lists the models that will be reported on in this research report. These were developed with two main applications in mind. The first application is to estimate the CRPS for daily precipitation forecasts produced by Bureau systems. The model 11QBS uses all possible predictors for this application, while 4QS and 7BS use subsets of these.

The second application is to estimate CRPS for 3-hourly precipitation forecasts produced by Bureau systems. The only issued forecasts for 3-hourly validity periods that can be considered are probability of no rain and quantile forecasts at the 0.5, 0.75 and 0.9 levels. The training data for this application uses daily forecasts, since CRPS can be reasonably calculated as per Section 3, but where forecasts with higher daily rainfall amounts are excluded. Specifically, since the highest mean 3-hourly precipitation predicted by AutoFcst was 30.4 mm, training data excludes all cases where predicted daily mean rainfall amount exceeds 30.4 mm. The 3QS and 4QBS models are used for this application.

Table 2: The predictors used by each linear model to predict CRPS.

Model name	**Predictors**
4QS	Quantile scores for quantile forecasts at the 0.25, 0.5, 0.75 and 0.9 levels
7BS	Brier scores for probability of non-exceedance forecasts at thresholds 0mm, 1mm, 5mm, 10mm, 15mm, 25mm and 50mm
11QBS	Union of predictands from 4QS and 7BS
3QS	Quantile scores for quantile forecasts at the 0.5, 0.75 and 0.9 levels
4QBS	Predictands from 3QB and Brier score for probability of no rain

5. Results

5.1. Model 4QS

Model 4QS, where all four quantile scores were used to predict the CRPS, gave exceptional results when trained and validated against the entire AutoFcst dataset with $R^2 = 0.9993$ and a root mean squared error (RMSE) in the prediction of CRPS of 0.169 mm. Table 3 gives the model coefficients, standard errors for those coefficients and the *F*-statistic for the model, which is significant at the 5% level.

Figure 2 shows the fit of 1000 randomly selected cases using model 4QS trained on the entire AutoFcst dataset, while Figure 3 shows the same plot zoomed into the bottom left corner.

Table 3: Coefficients, with standard errors in parentheses, for models 4QS and 3QS, along with *F*-statistics and R^2 values. Each *F*-statistic is significant at the 5% level.

Model	Intercept	$QS_{0.25}$	$QS_{0.5}$	$QS_{0.75}$	$QS_{0.9}$	*F*-stat	R^2
4QS	n/a	0.5234 (<0.0001)	0.5435 (<0.0001)	0.3461 (<0.0001)	0.3304 (<0.0001)	9.8×10^8	0.9993
3QS	n/a	n/a	0.9073 (<0.0001)	0.2699 (<0.0001)	0.3391 (<0.0001)	6.6×10^8	0.9986

By way of further illustration, the observation corresponding to the reconstructed predictive distribution of Figure 1 is 50.2 mm and its calculated CRPS is 16.08 mm. The estimated CRPS using model 4QS is 14.97 mm.

Tests showed that this model is not sensitive to overfitting. For example, when the model was trained using lead day 1 AutoFcst forecasts for stations in Tasmania for the 2021-22 summer, and then validated against AutoFcst for stations in the Australian tropics at various lead days and for various three-month periods, R^2 remained consistently above 0.999.

The model, trained on the full AutoFcst dataset, was also validated against Official lead day 1 forecasts for the year 2022, yielding $R^2 = 0.9987$ and $\mathrm{RMSE} = 0.193$ mm.

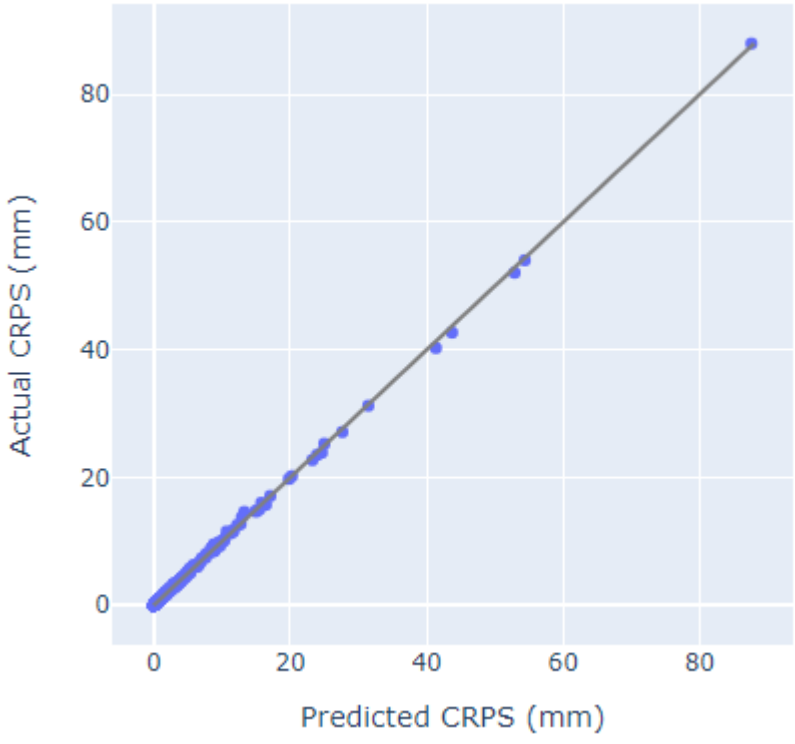

Figure 2: Predicted CRPS using model 4QS trained on the complete AutoFcst dataset.

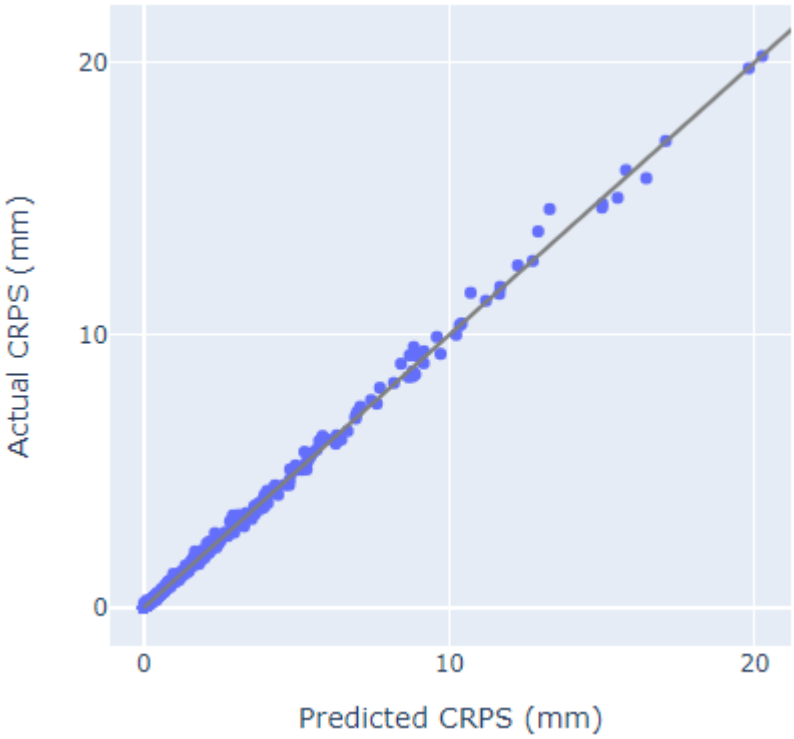

Figure 3: Predicted CRPS using model 4QS trained on the complete AutoFcst dataset as in Figure 2, but zoomed in.

5.2. Model 7BS

Model 7BS, where all seven Brier scores are used to predict CRPS, had $R^2 = 0.7570$ and $\mathrm{RMSE} = 3.11$ mm when trained and validated on the full AutoFcst dataset. This performance is mediocre relative to models using three or four quantile scores. Model coefficients, with standard errors, are presented in Table 4. The F-statistics for this model is 1.2×10^6, which is significant at the 5% level.

Tests showed that this model is reasonably insensitive to overfitting. For example, when the model was developed using lead day 1 AutoFcst forecasts for stations in Tasmania for the 2021-22 summer, and then validated against AutoFcst for stations in the Australian tropics at various lead days and for various three-month periods, R^2 remained consistently above 0.7.

Table 4: Coefficients, with standard errors in parentheses, for model 7BS.

Intercept	$\mathbf{BS_0}$	$\mathbf{BS_1}$	$\mathbf{BS_5}$	$\mathbf{BS_{10}}$	$\mathbf{BS_{15}}$	$\mathbf{BS_{25}}$	$\mathbf{BS_{50}}$
n/a	0.3439	2.2396	4.6657	5.1052	7.3031	14.7789	45.6709
	(0.009)	(0.013)	(0.017)	(0.022)	(0.025)	(0.024)	(0.032)

The model, trained on the full set of AutoFcst data, was also validated against Official lead day 1 forecasts for the year 2022, yielding $R^2 = 0.6982$ and $\mathrm{RMSE} = 2.90$ mm.

5.3. Model 11QBS

The model that uses all eleven quantile and probability exceedance forecasts did not perform substantially better than the model that uses only four quantile forecasts (4QS). The R^2 values for both 11QBS and 4QS models were equal when rounded to 4 decimal places, and the RMSEs were equal when rounded to 3 decimal places. Coefficients for probability of exceedance forecasts were non-zero only for the 0 mm and 1 mm non-exceedance thresholds, meaning that inclusion of probability of exceedance forecasts for higher thresholds added no skill to the model.

5.4. Model 3QS

As mentioned in Section 4, model 3QS is designed specifically for use for 3-hourly precipitation forecasts and as such forecast cases where mean precipitation amount exceeded 30.4 mm were excluded from all training and validation datasets.

When trained and validated against the AutoFcst dataset, the model gave very good results with $R^2 = 0.9986$ and an RMSE in the prediction of CRPS of 0.228 mm. Model coefficients, standard errors and the model *F*-statistic are presented in the final row of Table 3. when trained against AutoFcst are given by the equation below:

$$\mathrm{CRPS} \approx 0.907\ \mathrm{QS}_{0.5} + 0.270\ \mathrm{QS}_{0.75} + 0.339\ \mathrm{QS}_{0.9}.$$

Further tests showed that this model is not sensitive to overfitting. For example, when the model was developed using lead day 1 AutoFcst forecasts for stations in Tasmania for the 2021-22 summer, and then validated against AutoFcst for stations in the Australian tropics at various lead days and for various three-month periods, R^2 remained consistently above 0.998.

The model, trained on AutoFcst data, was also validated against Official lead day 1 forecasts for the year 2022, yielding $R^2 = 0.9970$ and $\mathrm{RMSE} = 0.229\ \mathrm{mm}$.

5.5. Model 4QBS

This model uses the three quantile forecasts of 3QS with the addition of the forecast probability of dry conditions. Using the same AutoFcst training dataset as 3QS, the coefficient for the Brier score for probability of dry conditions was 0. Consequently, this model gives the same CRPS predictions as model 3QS.

5.6. Discussion

It is interesting that models 4QS and 3QS have far superior performance to model 7BS. This is most likely because, on average, quantile forecasts at three or four suitably chosen but fixed quantile levels give more information about the any given predictive CDF than probably of non-exceedance forecasts for seven fixed thresholds. That is, the information about the predictive CDF from the four quantile forecasts (at quantile levels 0.25, 0.5, 0.75 and 0.9) is generally of high quality irrespective of the shape of the CDF. On the other hand, probability of non-exceedance forecasts at the

seven the fixed thresholds (0 mm, 1 mm, …, 50 mm) may miss some key features of some CDFs (e.g., if the bulk of the density lies beyond 50 mm) while containing much redundant information in other CDFs (e.g., if the probability of exceeding 5 mm is 0%).

6. Summary

The results of this report illustrate that the CRPS for operational daily precipitation predictive distributions can be estimated with a very high degree of accuracy using a weighted sum of four quantile scores via the formula

$$\mathrm{CRPS} \approx 0.523\ \mathrm{QS}_{0.25} + 0.543\ \mathrm{QS}_{0.5} + 0.346\ \mathrm{QS}_{0.75} + 0.330\ \mathrm{QS}_{0.9}\,. \qquad (5)$$

For operational 3-hourly precipitation forecasts, the approximation

$$\mathrm{CRPS} \approx 0.907\ \mathrm{QS}_{0.5} + 0.270\ \mathrm{QS}_{0.75} + 0.339\ \mathrm{QS}_{0.9} \qquad (6)$$

also has a high degree of accuracy. The scores given by either side of Equations (5) and (6) have same units as the observations. The quantile scores in both equations can be readily calculated using observations and published Bureau forecasts.

This study showed that similar approximations to the CRPS based on weighted sums of Brier scores do not approximate CRPS as accurately as those based on quantile scores, even though they are based on probability of exceedance forecasts for seven different thresholds. This is because the information about the predictive CDF contained in the quantile forecasts is richer than that contained in the probability of non-exceedance forecasts.

It is worth emphasising that the right-hand sides of Equations (5) and (6) are strictly proper scoring rules, irrespective of how close their values are to the CRPS. Given that they typically predict CRPS with a high degree of accuracy aids the interpretation of these scoring functions: they give roughly equal weight to the predictive performance of the forecast CDF for each quantile level α between 0 and 1, and roughly equal weight for each binary decision threshold in the range of possible outcomes.

Acknowledgments. The author would like to thank Deryn Griffiths for a fruitful discussion about ideas that lead to this report. Many thanks also to Tom Pagano, Morwenna Griffiths and Jason West for constructive feedback on an earlier version of this manuscript.

References

Gneiting, Tilmann, and Adrian E. Raftery. "Strictly proper scoring rules, prediction, and estimation." *Journal of the American statistical Association* 102, no. 477 (2007): 359-378.

Gneiting, Tilmann, and Roopesh Ranjan. "Comparing density forecasts using threshold-and quantile-weighted scoring rules." *Journal of Business & Economic Statistics* 29, no. 3 (2011): 411-422.

Laio, Francesco, and Stefania Tamea. "Verification tools for probabilistic forecasts of continuous hydrological variables." *Hydrology and Earth System Sciences* 11, no. 4 (2007): 1267-1277.

Matheson, James E., and Robert L. Winkler. "Scoring rules for continuous probability distributions." *Management science* 22, no. 10 (1976): 1087-1096.