Australian Government
Bureau of Meteorology

Assessing calibration when predictive distributions have discontinuities

Robert Taggart

June 2022

Bureau Research Report - 064

Assessing calibration when predictive distributions have discontinuities

Robert Taggart

Bureau Research Report No. 064

June 2022

National Library of Australia Cataloguing-in-Publication entry

Authors: Robert Taggart

Title: Assessing calibration when predictive distributions have discontinuities

ISBN: 978-1-925738-53-7
ISSN: 2206-3366

Series: Bureau Research Report – BRR064

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

It is desirable that predictive distributions generated by a forecast system exhibit good probabilistic calibration. The probability integral transform (PIT) is a well-established tool for measuring the probabilistic calibration when the predictive cumulative distribution functions (CDFs) are continuous, or equivalently when the distributions have density functions. However, predictive CDFs with at least one point of discontinuity are common in many situations, such as for precipitation forecasts at 0 mm or for streamflow forecasts at 0 $m^3 s^{-1}$ in drier climates.

Extensions of the classical definition of the PIT by Wang and Robertson (2011) in the empirical setting and by Gneiting and Ranjan (2013) in the prediction space setting allow probabilistic calibration to be measured when predictive CDFs are (possibly) discontinuous. In the extant literature, the application of these extensions involves taking random draws from the uniform distribution. In this report, an alternative method of applying the extention of the PIT is presented. Instead of taking random draws, the PIT of each forecast–observation pair is interpreted as a CDF, from which the empirical distribution of 'PIT values' from the joint sample of forecasts and observations is constructed. This empirical distribution can then be used to generate the usual visual tools for assessing mis-calibration, such as PIT diagrams (also known as PIT uniform probability plots) and PIT histograms (also known as rank histograms and Talagrand diagrams).

Finally, a novel statistic that measures calibration is presented, which can be decomposed into contributions primarily attributable to over/under-prediction and over/under-dispersion in the forecast system.

The main ideas of this report are illustrated using simple synthetic forecasts and operational precipitation forecasts issued by the Bureau of Meteorology. Software for computing probabalistic calabration using this approach has been implemented in the Bureau's `jive.metrics` package.

1 Introduction

Over the last decade or so, compelling arguments have been made that forecasts, where possible, ought to be probabilistic in nature (Gneiting and Katzfuss, 2014) and that predictive distributions should be as sharp as possible, subject to calibration (Gneiting et al., 2007; Gneiting and Ranjan, 2013). A forecast system is probabilistically calibrated if random draws from its predictive distributions are statistically indistinguishable from the corresponding observations. A consequence is that when a forecast system is probabilistically calibrated the observations will, for example, fall in the upper decile of its predictive distributions about 10% of the time. Essentially, the forecast system is reliable.

A key tool for assessing the calibration of predictive distributions is the probability integral transform (PIT) (Dawid, 1984; Diebold et al., 1998). Given a predictive cumulative distribution function (CDF) F and corresponding observation y, the associated PIT value (in its classical formulation) is the value $F(y)$. Well known results of Rosenblatt (1952) imply that if the CDFs are continuous (or equivalently, if the distributions have probability density functions), then the population of PIT values from a perfectly calibrated forecast system will be uniformly distributed. Probabilistic calibration is often assessed by visually inspecting any deviations from uniformity using so-called PIT diagrams and PIT histograms (Raftery et al., 2005), the latter being more or less the same as the rank histogram or Talagrand diagram which arose from assessing the calibration of ensemble forecasts (Anderson, 1996; Hamill and Colucci, 1997; Talagrand, 1999). These tools and concepts are reviewed and illustrated in Section 2, along with a graphical method of exposing conditional mis-calibration.

However, the classical formulation of the PIT cannot be used to identify probabilistic mis-calibration if any of the predictive CDFs are discontinuous. Such situations naturally arise when forecasting the accumulated precipitation over a short duration, because the forecast probability of observing exactly 0 mm is frequently positive. A similar problem arises for wind speed forecasts when the probability of calm conditions exceeds zero, or for streamflow forecasts for ephemeral streams. In slightly different contexts, Wang and Robertson (2011) and Gneiting and Ranjan (2013) gave an extension of the PIT that could accommodate CDFs with discontinuities. When applied in practice in the streamflow literature (e.g. Wang and Robertson (2011), Bennett et al. (2017)), whenever an observation coincides with a point of discontinuity of the predictive CDF, a random draw is taken from a suitable uniform distribution to represent the PIT value for that for forecast–observation pair. This adds some instability to the summary statistics calculated from the resulting sample of PIT values. Stabilization is possible by repeating the process many times and calculating the mean of the resulting statistics. However, this report presents an approach that avoids random draws altogether.

Classically, the PIT value for a forecast–observation pair is a number from the unit interval $[0, 1]$. The approach taken in this in this report is to interpret the extension of the PIT for a forecast–observation pair as a CDF on the unit interval $[0, 1]$. Section 3 explains how this works from a practitioner's viewpoint and illustrates how it can be used to assess the probabilistic calibration of a forecast system where some of the predictive CDFs are discontinuous. The key idea is that the empirical distribution of PIT values for the sample can be calculated by taking the pointwise arithmetic mean of the PIT values, interpreted as CDFs. It is then shown how this empirical distribution is used to construct

the usual visual tools for assessing probabilistic calibration.

In Section 5 a novel statistic for measuring calibration is introduced, which can be decomposed into components that signal contributions from under/over-prediction and from under/over-dispersion of the forecast system. This provides an alternative measure of calibration to the so-called α-index (Renard et al., 2010), which has gained some popularity in the streamflow and seasonal forecast literature (e.g. Bennett et al. (2017); Berthet et al. (2020); Li et al. (2020); Shao et al. (2021)).

All concepts in this paper are illustrated using simple synthetic examples and operational forecasts for precipitation issued by the Australian Bureau of Meteorology (BoM). Software for computing probabalistic calibration using the approach in this report has been implemented in the Bureau's `jive.metrics` Python package.

2 Calibration and the classical PIT

The rise of ensemble prediction systems over the last few decades has provided a fruitful source from which to construct predictive distributions for many weather variables, based on the distribution of forecast values taken from the ensemble members. Assessing whether such distributions are well-calibrated is a critical step to evaluating whether the ensemble captures the appropriate range of possible realizations, and to what extent statistical post-processing of the predictive distributions is required.

In the context of predictive distributions for a real-valued variable, Gneiting et al. (2007) discuss three notions of calibration (probabilistic calibration, exceedance calibration and marginal calibration) and provide simple examples to demonstrate that these notions are distinct. In addition, Gneiting and Ranjan (2013) also compare these with the notion of auto-calibration, and conclude 'Generally, probabilistic calibration continues to be the most practically useful and most practically relevant notion of calibration,' while in settings 'such as climate prediction, ideas closely related to marginal calibration play crucial roles.' This report focuses on probabilistic calibration.

A predictive distribution for some random variable Y that takes values in the real line $\mathbb{R}$ can be represented as a CDF F, so that $\mathbb{P}(Y \leq x) = F(x)$ whenever $x \in \mathbb{R}$. If Y has a distribution given by F then we write $Y \sim F$. If F is continuous (as it typically the case for temperature forecasts) then the predictive distribution can also be represented as a probability density function. But not all predictive CDFs are continuous. For example, if Y is daily precipitation and the forecast chance of no precipitation is 40%, then the left-hand limit at 0 mm satisfies $\lim_{s \uparrow 0} F(s) = 0 \neq F(0) = 0.4$ and hence F is not continuous at 0.

Probabilistic calibration concerns the joint distribution of forecasts and observations and is defined using the probability integral transform (PIT). If $Y \sim F$ then the PIT of Y is classically defined as $F(Y)$. Note that $0 \leq F(Y) \leq 1$. A standard result is that if F is continuous and $Y \sim F$ then the PIT $F(Y)$ is a standard uniform random variable (Rosenblatt, 1952; Pearson, 1933). A forecast system that issues continuous predictive CDFs is *probabilistically calibrated* if its population of PIT values has standard uniform distribution. In practice within an empirical setting, one takes a sample $\{(F_i, y_i)\}_{i=1}^n$ of the joint distribution of continuous predictive CDFs F_i and corresponding real-valued observations y_i. The ith PIT value q_i from the sample is given by $q_i = F_i(y_i)$, and the distribution of sampled PIT values $\{q_i\}_{i=1}^n$ is compared with the standard uniform

distribution.

Noting that the standard uniform distribution has an expected value of 1/2 and a variance of 1/12, we say that a sample of predictive distributions

- exhibits an *over-prediction tendency* if the sample mean of the PIT values is less than 1/2,
- exhibits an *under-prediction tendency* if the sample mean of the PIT values is greater than 1/2,
- is *over-dispersed* if the sample variance of the PIT values is less than 1/12, and
- is *under-dispersed* if the sample variance of the PIT values is greater than 1/12.

Over-dispersion occurs when the predictive densities are too wide (i.e., too many observations falling in the center of the predictive distributions) while under-dispersion occurs when the predictive densities are too narrow (i.e., too many observations falling in the tails of the predictive distributions). The definitions of probabilistic calibration and over/under-dispersion presented within this emprical setting are 'practitioner's analogues' of those presented by Gneiting and Ranjan (2013), which use the measure-theoretic notion of a prediction space.

Standard graphical devices for identifying any probabilistic mis-calibration include the PIT histogram and PIT diagram. PIT histograms show the frequencies of binned PIT values. PIT diagrams, also called 'predictive QQ plots' (Thyer et al., 2009) or 'PIT uniform probability plots' (Wang et al., 2009), are either graphs of the empirical CDF (eCDF) of PIT values, or plots the points $\{(q_{(i)}, i/(n+1))\}_{i=1}^{n}$, where $q_{(i)}$ denotes the ith smallest PIT value from the sample. As noted by several authors (e.g. Laio and Tamea (2007), Gneiting et al. (2007)) and illustrated by the following example, different types of mis-calibration result in different characteristic shapes in these histograms or diagrams.

Example 2.1. Consider a process that generates, for each time step t, an observation sampled from the normal distribution $\mathcal{N}(\mu_t, 1)$ with mean μ_t and variance 1, where $\mu_t \sim \mathcal{N}(0, 1)$. Using 100,000 forecast cases with independently generated observations, Figure 1 illustrates the characteristic shapes of PIT histograms and PIT diagrams for forecast systems that produce different types of predictive distributions F_t.

- Ideal (and thus well-calibrated) predictive distributions $F_t = \mathcal{N}(\mu_t, 1)$ generate a flat histogram and eCDF lying on the diagonal.
- Predictive distributions $F_t = \mathcal{N}(\mu_t + 0.25, 1)$ with an over-prediction tendency generate a decreasing histogram and eCDF lying 'over' the diagonal.
- Predictive distributions $F_t = \mathcal{N}(\mu_t - 0.25, 1)$ with an under-prediction tendency generate an increasing histogram and eCDF lying 'under' the diagonal;
- Predictive distributions $F_t = \mathcal{N}(\mu_t, 1.2^2)$ that are over-dispersed generate an inverted 'U' shaped histogram and eCDF lying 'under–over' with respect to the diagonal.
- Predictive distributions $F_t = \mathcal{N}(\mu_t, 1.2^{-2})$ that are under-dispersed generate a 'U' shaped histogram and eCDF lying 'over–under' with respect to the diagonal.

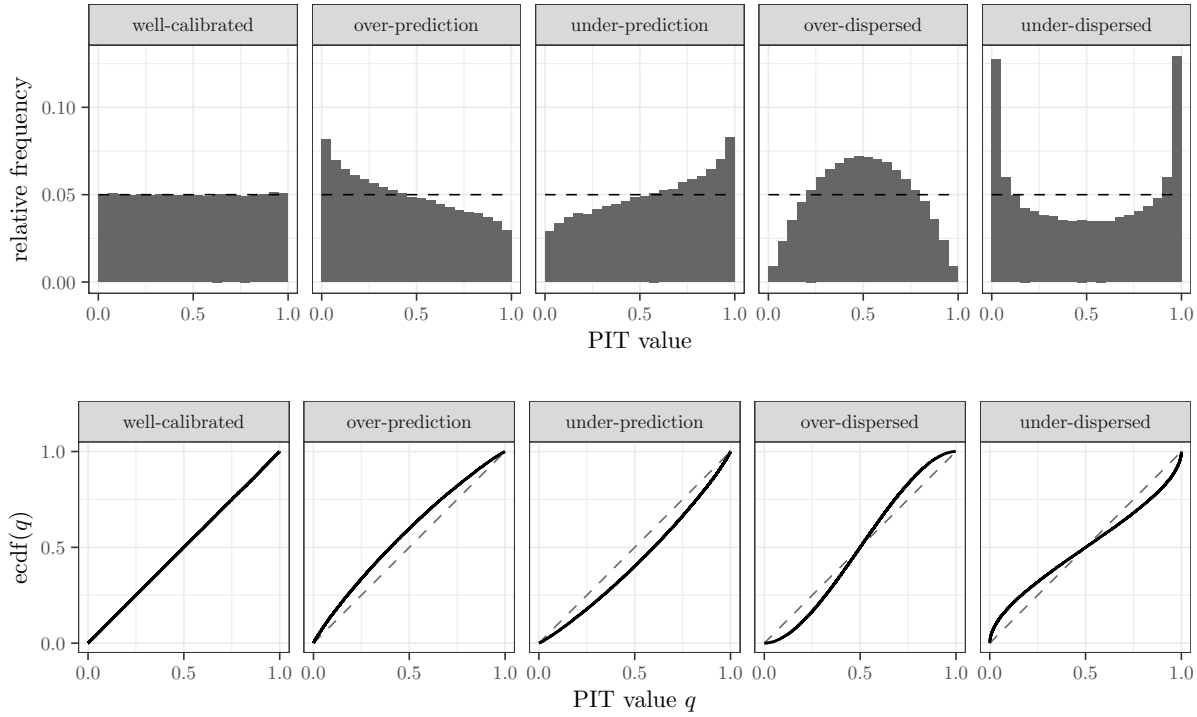

Figure 1: Characteristic shapes of PIT histograms (top) and PIT diagrams (bottom) for forecast systems that are well-calibrated, have over-prediction tendency, have under-prediction tendency, are over-dispersed and are under-dispersed. Each PIT histogram has 20 equally spaced bins. Each PIT diagram is the graph of the eCDF of the sample PIT values q_i.

Finally, if the predictive distributions F_t are based on climatology (i.e., $F_t = \mathcal{N}(0, 2)$) then the corresponding PIT histogram is also flat (i.e., essentially the same as the 'well-calibrated' PIT histogram in Figure 1) and the forecast is probabilistically calibrated. However the resulting predictive CDFs are not sharp and hence the climatology-based forecast system will be less useful than a forecast system with slight bias but substantially sharper predictive distributions. Superior probabilistic calibration needn't imply a better forecast.

As noted by Hamill (2001), care must be taken when interpreting PIT histograms. A flat PIT histogram is a necessary but not sufficient characteristic of forecasts from a well-calibrated system. A system may have significant conditional mis-calibration but may generate a flat PIT histogram, as illustrated by the following example.

Example 2.2. Hamill (2001) Consider a process that, for each time step t, generates an observation y_t sampled from $\mathcal{N}(0, 1)$. Suppose that, for each t, the forecaster's predictive distribution F_t is, with equal likelihood, one of $\mathcal{N}(-0.5, 1)$ (under-prediction bias), $\mathcal{N}(0.5, 1)$ (over-prediction bias) or $\mathcal{N}(0, (1.3)^2)$ (over-dispersed). A PIT histogram, generated from sample of size 100,000 forecast cases and illustrated in the left-hand panel of Figure 2, is near flat, masking the fact that the forecast system has significant conditional mis-calibration.

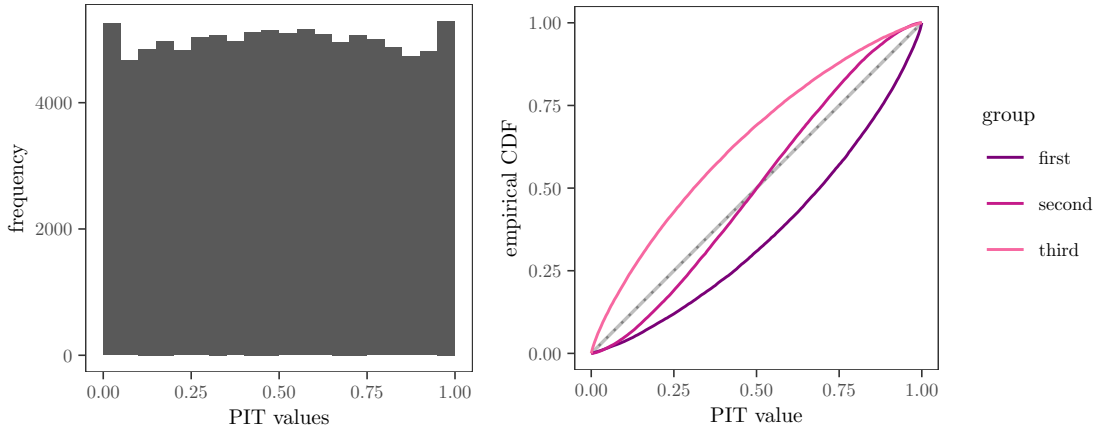

Figure 2: The near-flat PIT histogram (left) in the synthetic forecast example of Hamill masks the fact that the system has substantial conditional mis-calibration, as illustrated by the graphs of eCDFs of PIT values (right), where the partition of forecast cases is based on mean values of the predictive distributions. In the plot on the right hand side, the dashed line is the CDF of the standard uniform distribution and the very narrow gray diagonal band shows the bounds of a Kolmogorov–Smirnoff significance test for uniformity at the 5% level.

Conditional mis-calibration can be checked by partitioning the set of forecast cases into a suitable number of bins, and plotting the graph of the eCDF of PIT values for each bin on the same diagram. The right-hand panel of Figure 2 shows the eCDFs of PIT values for Hamill's synthetic example, where the forecast cases are partitioned into three groups based on the expected (i.e. mean) values of the predictive distributions. The first group (predictive distributions with an expected value of -0.5) consists of predictive distributions with an under-prediction tendency, as evidenced by the corresponding curve lying under the dashed diagonal line. Similarly, the diagram shows that the second group (i.e., predictive distributions with an expected value of 0) consists of over-dispersed predictive CDFs, while the third group has an over-prediction tendency. The very narrow light gray region about the diagonal shows the bounds of a Kolmogorov–Smirnoff significance test for uniformity at the 5% level. The eCDFs of PIT values of all three groups stray well outside this region, so the null hypothesis of probabilistic calibration can be rejected at the 5% significance level for each group.

Such partitions induce decompositions of the distribution of PIT values, since the eCDF of all PIT values is equal to a weighted mean of the eCDFs of the grouped PIT values, with the weightings determined by the number of samples in each group.

3 Extension of the PIT as a CDF-valued operator

In the context of prediction spaces, Gneiting and Ranjan (2013) gave an extension of the PIT by reformulating it as a CDF-valued random quantity. This extension can be used to assess the probabilistic calibration of a forecast system even if its predictive CDFs have points of discontinuity. In the practitioners' empirical setting, Wang and Robertson (2011)

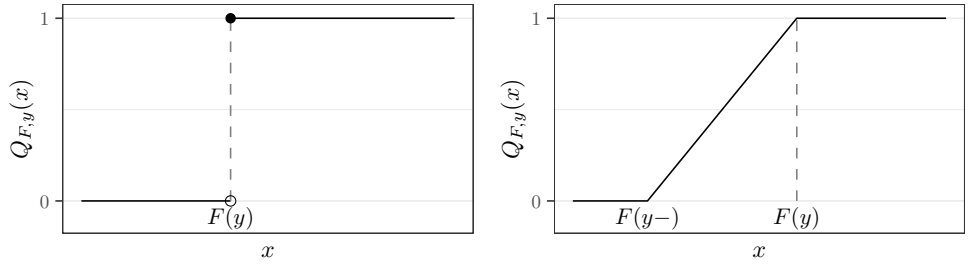

Figure 3: The PIT $Q_{F,y}$ when F is continuous at the observation y (left panel) and when F is discontinuous at the observation y (right panel).

extend the PIT so that it is able to handle discontinuities using somewhat similar ideas, without explicitly reformulating the PIT as a CDF-valued operator. Inspired by these ideas, we define the PIT as a CDF-valued operator in a way that is useful for practitioners in an empirical setting.

Suppose that F is a predictive CDF and y the corresponding observation. Then the CDF-valued PIT for the pair (F, y), denoted here by $Q_{F,y}$, is defined by

$$Q_{F,y}(x) = \begin{cases} 0, & x \leq F(y-), \\ \dfrac{x - F(y-)}{F(y) - F(y-)}, & F(y-) < x < F(y), \\ 1, & x \geq F(y), \end{cases} \tag{3.1}$$

whenever $x \in \mathbb{R}$. Here $F(y-)$ denotes the left-hand limit $\lim_{s \uparrow y} F(s)$.

To see what this means, consider two cases. In the case when F is continuous at y, $F(y-) = F(y)$ and hence the PIT has the 'classical PIT value' $F(y)$ with probability 1, and its graph is illustrated in the left-hand panel of Figure 3. In the case when F is discontinuous at y, the graph of $Q_{F,y}$ is illustrated in the right-hand panel of Figure 3 and represents the CDF of a random variable with uniform distribution on the interval $[F(y-), F(y)]$.

Given a finite sample $\{(F_i, y_i)\}_{i=1}^n$ of forecast–observation pairs from a forecast system, the corresponding eCDF of PIT values, denoted by $\bar{Q}$, is given by

$$\bar{Q}(x) = \frac{1}{n} \sum_{i=1}^{n} Q_{F_i, y_i}(x) \tag{3.2}$$

whenever $x \in \mathbb{R}$. We shall also refer to $\bar{Q}$ as the *PIT eCDF* of the sample. This formulation of the empirical distribution of PIT values agrees with the eCDF derived using classical PIT values in the case where each predictive distribution F_i is continuous.

The notions introduced in Section 2 of over- and under-prediction, and of over- and under-dispersion, apply in this context by calculating the expected value and variance of the eCDF $\bar{Q}$. The sample has probabilistic mis-calibration to the extent that $\bar{Q}$ deviates from the standard uniform distribution. This can be assessed by plotting the graphs of $\bar{Q}$ and the standard uniform CDF on the same diagram. PIT histograms can be constructed from $\bar{Q}$ as follows: if a bin in the desired histogram is the interval $(a, b]$ then the 'relative frequency of PITs' for that bin is given by $\bar{Q}(b) - \bar{Q}(a)$. These applictions are illustrated in the following three sections.

4 The CDF-valued PIT: Examples

Example 4.1. Consider the hybrid exponential distribution with CDF F given by

$$F(s) = \begin{cases} 0, & s < 0 \\ 0.35 + 0.65\left(1 - e^{-s/5}\right), & s \geq 0. \end{cases}$$

The graph of F is shown in the top panel of Figure 4. Note that F is discontinuous at 0. Interpreted as a forecast for daily precipitation, F gives a 35% chance of no rain, but if rain does occur then its accumulated value has an exponential distribution. Consider n forecast–observation pairs $(F, y_i)_{i=1}^n$, where each observation y_i is an independent random draw from the distribution F. Figure 4 shows the graphs of $\bar{Q}$ for different sample sizes n. For the sample size of 10, the observation 0 was drawn twice, which is less than the expected frequency of 3.5. As a consequence, the graph of $\bar{Q}$ starts with a sloped segment that lies below the dashed diagonal line. The remainder of the graph is piecewise horizontal with step jumps at each $F(y_i)$ for which $y_i \neq 0$. The gray shaded region in the plot is the region associated with a Kolmogorov–Smirnoff significance test of 5%. This is the region in which the graph of an eCDF, constructed from n independent random draws from the standard uniform distribution, would lie 95% of the time. Naturally, as the sample size n increases, the shaded region narrows and $\bar{Q}$ lies closer to the diagonal.

Example 4.2. Predictive distributions for accumulated daily precipitation are issued by the BoM across a gridded Australian domain. We consider lead day 1 forecasts at two locations (Sydney Airport and Perisher Valley) in New South Wales, Australia, for the three year period July 2018 to June 2021. The left and center panels of Figure 5 show graphs of the eCDF $\bar{Q}$ and PIT histograms with 10 bins of equal width. Forecasts for both locations show slight to moderate under-forecast bias coupled with some over-dispersion. The light gray shaded region in the left panel shows the bounds of a Kolmogorov–Smirnoff significance test for uniformity at the 5% level. The graph of $\bar{Q}$ for Perisher Valley clearly falls outside this region and therefore the null hypothesis that Perisher Valley forecasts are probabilistically calibrated can be rejected.

The right panel of Figure 5 shows a graphical check for conditional mis-calibration based on the expected precipitation. That is, forecasts have been grouped into four equal bins based on quartile cuts of the distribution of the expectations (i.e. mean value) of the predictive distributions, and $\bar{Q}$ calculated for each group. The lines with the two lightest hues show the graphs of $\bar{Q}$ for the two lowest quartiles, and almost coincide for much their range. The line with the darkest hue shows the graph of $\bar{Q}$ for group 4, which is the highest quartile. At Perisher Valley, group 4 exhibits strong under-prediction tendency and some over-dispersion. Other groups with lower expected precipitation at this location have PIT eCDFs that are close to uniform and fall within the 5% Kolmogorov–Smirnoff bands.

5 Statistics that measure probabilistic mis-calibration

It is convenient to have a statistic that measures probabilistic mis-calibration; that is, the extent to which the PIT eCDF $\bar{Q}$ deviates from the CDF V of the standard uniform

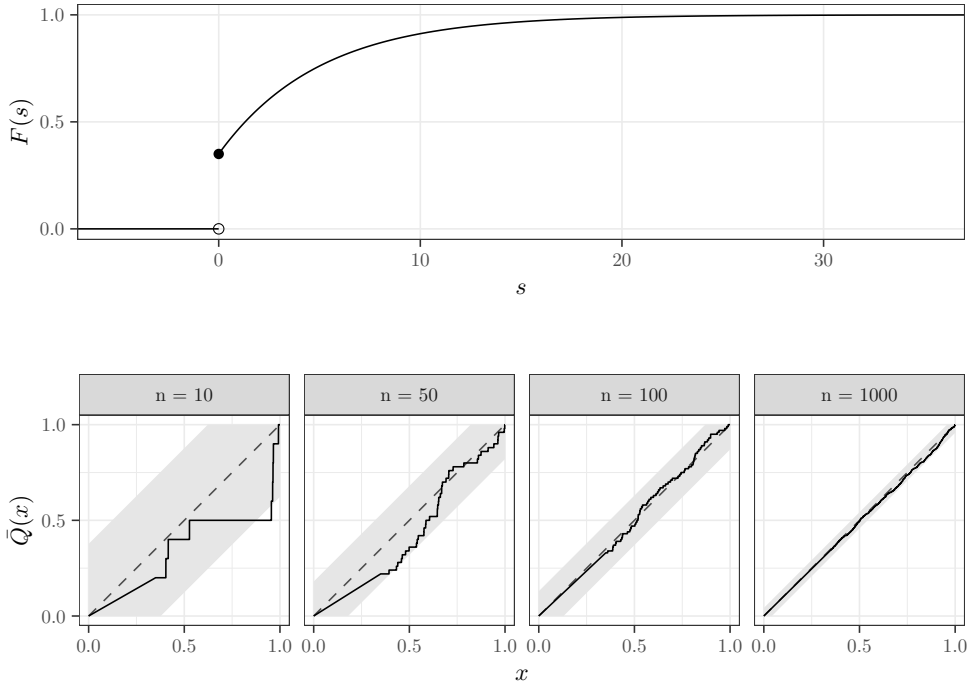

Figure 4: Top: The CDF F of a hybrid exponential function. Bottom: The PIT eCDF $\bar{Q}$ (solid black line) obtained from taking n random independent draws from F. The dashed gray diagonal line is the CDF of the standard uniform distribution. The light gray shadded region shows the bounds of a Kolmogorov–Smirnoff significance test for uniformity at the 5% level.

distribution. Recall that $V(x) = x$ on $[0, 1]$. Three options, identified using the label 'PIT Score' (PS), are

$$\mathrm{PS}_1(\bar{Q}) = \int_0^1 |\bar{Q}(x) - x| \, \mathrm{d}x, \tag{5.1}$$

$$\mathrm{PS}_2(\bar{Q}) = \int_0^1 |\bar{Q}(x) - x|^2 \, \mathrm{d}x \tag{5.2}$$

and

$$\mathrm{PS}_\infty(\bar{Q}) = \sup_{x \in [0,1]} |\bar{Q}(x) - x|. \tag{5.3}$$

The statistic $\mathrm{PS}_1(\bar{Q})$ is closely related to the α-index of Renard et al. (2010), which has found regular usage in the streamflow literature. $\mathrm{PS}_2(\bar{Q})$ is the test statistic for a Cramér–von Mises significance test for uniformity. $\mathrm{PS}_\infty(\bar{Q})$ is essentially the maximum vertical distance between the graph of the PIT eCDF and the diagonal line representing V, and is the test statistic of a Kolmogorov–Smirnoff significance test for uniformity. The remainder of this section focuses on PS_2, because it can be decomposed into terms related to under/over-prediction and under/over-dispersion.

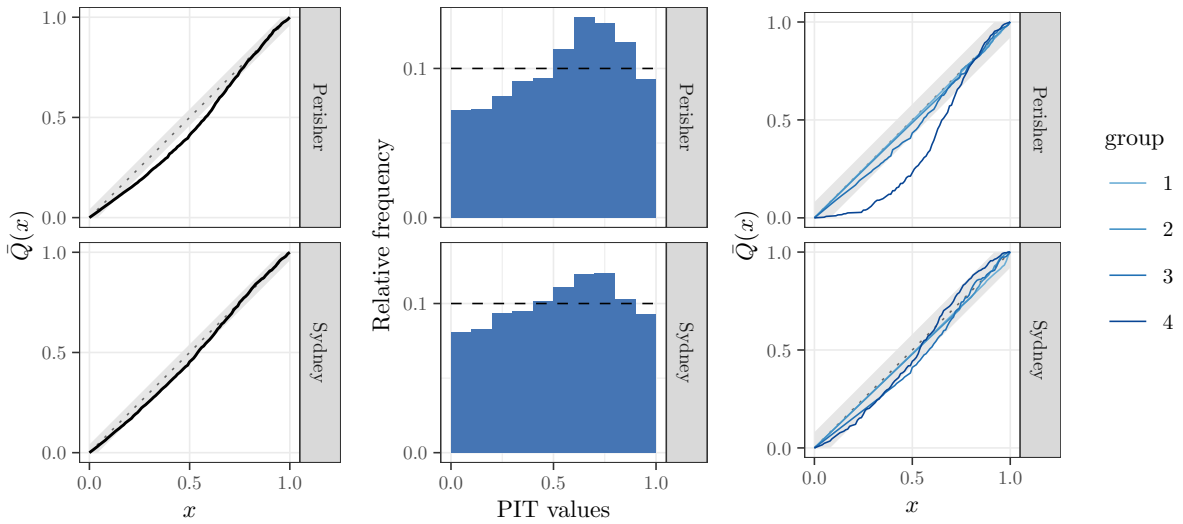

Figure 5: Calibration of BoM daily precipitation forecasts at Perisher Valley and Sydney Airport at lead day 1. Left: PIT diagram showing the graph of the eCDF $\bar{Q}$ of PIT values. Center: PIT histograms using 10 bins. Right: PIT diagram showing the PIT eCDF $\bar{Q}$ for each group of forecast cases, where forecasts are grouped by the expectation of their predictive distributions. Group 1 (lightest hue) consists of forecast cases with expected precipitation in the lowest quartile, while Group 4 (darkest hue) consists of forecast cases with expected precipitation in the highest quartile. The light gray shadded region (left and right) shows the bounds of a Kolmogorov–Smirnoff significance test for uniformity at the 5% level.

Proposition 5.1. *Suppose that $\bar{Q}$ is the PIT eCDF for some finite set of forecast–observation pairs. Let $\bar{Q}^{-1}$ denote the generalized inverse distribution function of $\bar{Q}$ and let V denote the CDF of the standard uniform distribution. Suppose that $X \sim \bar{Q}$ and $U \sim V$ with X and U independent. Then*

$$\mathrm{PS}_2(\bar{Q}) = (\mathbb{E}(X) - \mathbb{E}(U))^2 + \mathrm{Var}(\bar{Q}^{-1}(U) - U) \tag{5.4}$$

$$= (\mathbb{E}(X) - \mathbb{E}(U))^2 + (\mathrm{Var}(U) - \mathrm{Var}(X)) + 2\,\mathrm{Cov}(\bar{Q}^{-1}(U), \bar{Q}^{-1}(U) - U). \tag{5.5}$$

The proof is given in the appendix. The terms in each decomposition can be interpreted as follows. The expression $\mathbb{E}(X) - \mathbb{E}(U)$ is the difference between the expected value of the PIT eCDF $\bar{Q}$ with the expected value of the standard uniform distribution. Recalling that the latter equals $1/2$, $\mathbb{E}(X) - \mathbb{E}(U)$ measures whether the forecast system has under-prediction tendency ($\mathbb{E}(X) - \mathbb{E}(U) > 0$) or over-prediction tendency ($\mathbb{E}(X) - \mathbb{E}(U) < 0$). Hence the first term of the right-hand side in each of Equations (5.4) and (5.5) indicates the degree to which under- or over-prediction is a problem. As a convenient shorthand we call this the *bias component*.

The second term in the decomposition (5.4) can be interpreted as the degree of spread of calibration error. To see why, note that since U is uniform, $\bar{Q}^{-1}(U)$ has distribution $\bar{Q}$ (think of $\bar{Q}^{-1}$ as the quantile function for distribution of classical PIT values). Thus $\bar{Q}^{-1}(U)$ represents a randomly selected classical PIT value, U its ideal relative rank among the PIT values, $\bar{Q}^{-1}(U) - U$ the signed error for that classical PIT value and $\mathrm{Var}(\bar{Q}^{-1}(U) -$

U) the variance of those errors. For convenience we call last term of Equation (5.4) the *variance component*.

Overall, the decomposition of Equation (5.4) is analogous to the well-known decomposition of mean squared error as the sum of the bias squared and the variance of error.

The second term in Equation (5.5) measures the dispersive error of a forecast system. Recall that $\mathrm{Var}(U) = 1/12$ and that this value indicates neutral dispersion. If $\mathrm{Var}(U) - \mathrm{Var}(X) > 0$ then the system is over-dispersed while if $\mathrm{Var}(U) - \mathrm{Var}(X) < 0$ then it is under-dispersed.

The final term in Equation (5.5) can be interpreted as a measure of the correlation between the distribution of PIT values $\bar{Q}^{-1}(U)$ and the distribution of their signed errors $\bar{Q}^{-1}(U) - U$.

Overall, the decomposition of Equation (5.5) is analogous to the three-term decomposition of mean squared error of Murphy and Winkler (1987).

Key terms in the decompositions of Proposition 5.1 can be calculated using the fact that if $F : [0, 1] \to [0, 1]$ is a CDF and $Z \sim F$ then

$$\mathbb{E}(Z) = \int_0^1 (1 - F(z))\,\mathrm{d}z$$

and

$$\mathrm{Var}(Z) = \int_0^1 z(1 - F(z))\,\mathrm{d}z - (\mathbb{E}(Z))^2.$$

Figure 6 shows, for the five synthetic forecast systems of Example 2.1, the values of $\mathrm{PS}_2(\bar{Q})$ and their decomposition as per Equation (5.4), calculated using 100,000 independently sampled forecast–observation pairs. The values of the bias component $(\mathbb{E}(X) - \mathbb{E}(U))^2$ of the decomposition correctly identify the forecast systems with tendency towards under- or over-prediction. The variance component $\mathrm{Var}(\bar{Q}^{-1}(U) - U)$ is greatest in forecast systems that are over- or under-dispersed, but is still positive in neutrally dispersed forecast systems that exhibit under- or over-prediction. This is because any PIT eCDF with under/over-prediction tendency must also have positive variance in the PIT errors $\bar{Q}^{-1}(U) - U$, since $\bar{Q}$ must always be nondecreasing and satisfy $\bar{Q}(0) = 0$ and $\bar{Q}(1) = 1$. That is, variance in the errors can never be independent of under/over-prediction tendency. Nonetheless, in these examples the relative values of the bias and variance components of $\mathrm{PS}_2(\bar{Q})$ do give a clear signal of the primary source of mis-calibration.

The second decomposition of $\mathrm{PS}_2(\bar{Q})$, given by Equation (5.5), is harder to interpret in practice because the last two terms of the decomposition tend to be strongly correlated and may also be negative. This will be illustrated using BoM forecast data in Section 6.

It is not hard to show that the possible values of $\mathrm{PS}_2(\bar{Q})$ are constrained to the interval $[0, 1/3]$, with values for the bias component $(\mathbb{E}(X) - \mathbb{E}(U))^2$ constrained to $[0, 1/4]$, and values for the variance component $\mathrm{Var}(\bar{Q}^{-1}(U) - U)$ constrained to $[0, 1/12]$. For those who prefer positively oriented indices, the statistic $1 - 3\,\mathrm{PS}_2(\bar{Q})$ ranges from 0 (poor calibration) to 1 (perfect calibration).

As with the Kolmogorov–Smirnoff test for uniformity, what constitutes a good value for $\mathrm{PS}_2(\bar{Q})$ depends on the sample size of the joint forecast–observation distribution. Figure 7 gives critical values for $\mathrm{PS}_2(\bar{Q})$ at the 1%, 5% and 10% significance levels when using

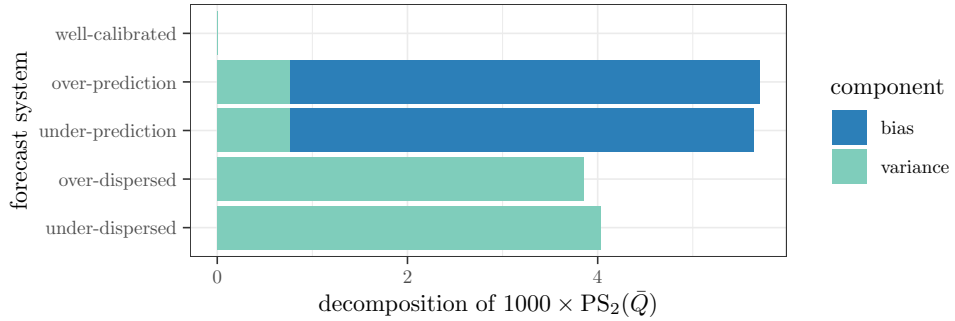

Figure 6: $\mathrm{PS}_2(\bar{Q})$ and its decomposition into the bias component $(\mathbb{E}(X) - \mathbb{E}(U))^2$ (darker hue) and variance component $\mathrm{Var}(\bar{Q}^{-1}(U) - U)$ (lighter hue) for the synthetic forecast systems of Example 2.1.

$\mathrm{PS}_2(\bar{Q})$ to reject the null hypothesis that $\bar{Q}$ was sampled from a uniform distribution. These critical values were calculated via a standard bootstrapping technique, where eCDFs G were constructed from randomly generated standard uniform samples, and $\mathrm{PS}_2(G)$ calculated. The standard error for each bootstrapped critical value estimate is at least three orders of magnitude smaller than the estimate. Figure 7 shows that the relationship between the log of the critical value and the log of the sample size is almost linear for the range of sample sizes computed. Note that the calculation of these critical values only used eCDFs that are step functions. When predictive CDFs are discontinuous the PIT eCDF will be a linear combination of step functions and piecewise linear continuous functions.

6 Application to BoM daily precipitation forecasts

We return to the daily precipitation forecasts for Perisher Valley and Sydney Airport to illustrate some of the ideas of Section 5. The top left panel of Figure 8 shows $\mathrm{PS}_2(\bar{Q})$ values for each location by lead day. Lower values indicate better calibrated forecasts. To get a sense of sampling error, 95% intervals have been calculated by bootstrapping the forecast–observation pairs (F_i, y_i) on the index i and taking the 0.025- and 0.975-quantiles of $\mathrm{PS}_2(\bar{Q})$ from the bootstrapped sample. The upper boundary of the gray region on the graph shows the critical value for the sample size at the 5% significance level. All $\mathrm{PS}_2(\bar{Q})$ values lie above this line. Assuming that, for each location and lead day, the set of PITs $\{Q_i\}_{i=1}^n$ are independent, one can reject the null hypothesis that the PIT eCDF is sampled from a uniform distribution at the 5% significance level. The error bars also lie above this line. Thus the forecasts are not probabilistically well calibrated.

The top right panel of Figure 8 shows the decomposition of $\mathrm{PS}_2(\bar{Q})$ into the bias $(\mathbb{E}(X) - \mathbb{E}(U))^2$ and variance $\mathrm{Var}(\bar{Q}^{-1}(U) - U)$ components. The bias component is the most significant contributor to mis-calibration and generally moves in step with $\mathrm{PS}_2(\bar{Q})$ across lead time. The variance component is relatively stable with lead time. The significant improvement in the calibration of Perisher Valley forecasts from lead days 3 to 1 is attributable primarily to a reduction in under-prediction tendency.

The center panels of Figure 8 show how under/over-prediction tendency ($\mathbb{E}(\bar{Q})$ relative to 0.5) and under/over-dispersion ($\mathrm{Var}(\bar{Q})$ relative to 1/12) vary with lead day. All fore-

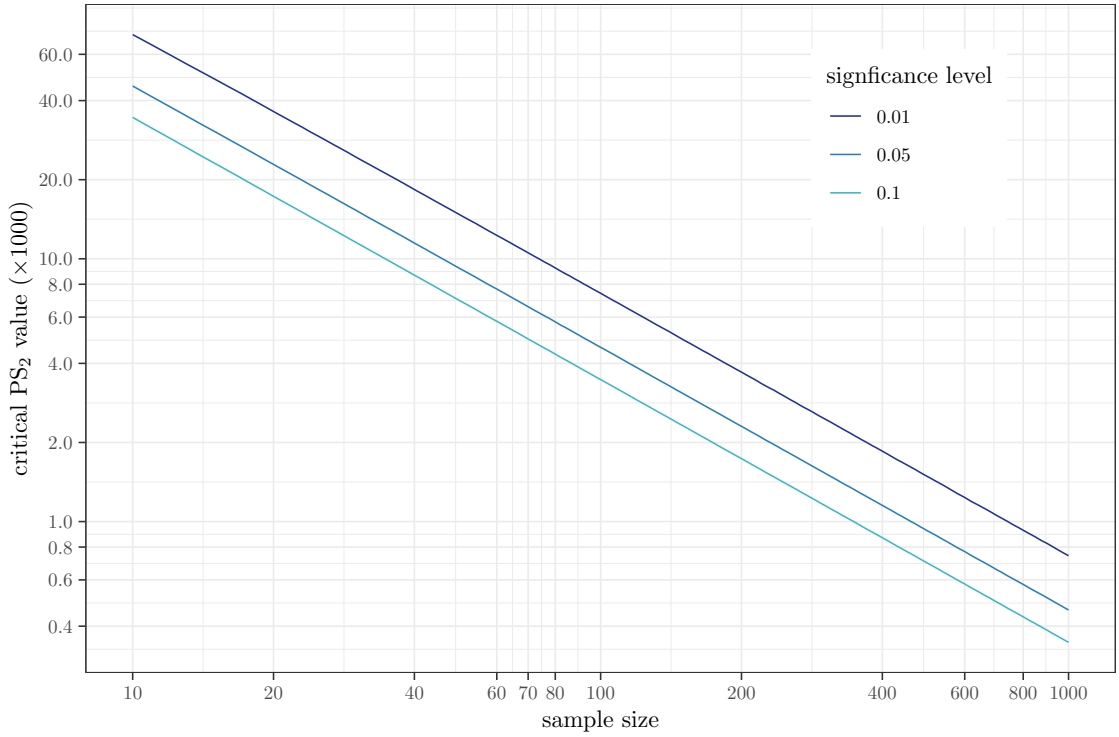

Figure 7: Critical values of PS_2 for the standard uniform distribution, multiplied by 1000. Both axes use a logarithmic scale.

casts have an under-prediction bias. Whilst Sydney Airport forecasts are over-dispersed, Perisher Valley forecasts switch from being under-dispersed at longer lead times to over-dispersed at short lead times. The change in calibration characteristics at Perisher Valley by lead day is primarily explained by different forecast production methods. At longer lead times the forecast is typically based on an automated consensus forecast system, which fails to resolve local orographic effects. At shorter lead times, professional meteorologists intervene to reduce the known under-prediction tendency of the automated system, and whilst doing so also change dispersion characteristics.

The bottom left panel of Figure 8 shows the dispersion error (dotted lines) and covariance (dashed lines) terms of the decomposition given by Equation (5.5). A key point, as illustrated here, is that in practice these terms are usually strongly negatively correlated. Of the panels in Figure 8 discussed so far, the most useful combination of information is gained from an overall measure of mis-calibration using $PS_2(\bar{Q})$ (via either one of the top panels), and the graphs showing under/over-prediction tendency and under/over-dispersion (centre panels).

The bottom right panel of Figure 8 shows the mean continuous ranked probability score (CRPS) (Matheson and Winkler, 1976) for each forecast system by lead day. The CRPS is a commonly used proper scoring rule for predictive distributions. When averaged over many forecast cases, the mean CRPS provides a good summary of the overall accuracy of a forecast system, with forecast systems producing sharp well-calibrated predictive distributions being rewarded. A lower CRPS is better. Note that all forecasts

become more accurate, as measured by mean CRPS, with decreasing lead time. This holds notwithstanding the tendency towards increasing mis-calibration from lead days 7 to 3. This highlights again that better calibration does not necessarily imply a better forecast, since the sharpness of predictive distributions also contributes to forecast accuracy. Moreover, when using large sample sizes, and systematic conditional mis-calibration in a forecast system will generally result in a higher mean CRPS, whereas summary calibration statistics derived from the entire set of forecast cases may not detect conditional mis-calibration (c.f. Example 2.2).

7 Discussion

The probability integral transform (PIT) has been a key tool for assessing the calibration of predictive distributions for several decades. In meteorological prediction, this has been driven by the increasing availability of ensemble prediction systems, from which one can construct predictive distributions. Visual diagnostic tools for assessing mis-calibration include PIT diagrams and PIT histograms, the latter of which are essentially the same as rank histograms and Talagrand diagrams. However, as classically defined, the PIT can only be used for assessing calibration when the predictive CDFs are continuous. Continuity typically holds for predictive distributions of temperature, but often fails for predictive distributions of precipitation, wind speed and streamflow for ephemeral streams. In this paper, we have used ideas of Wang and Robertson (2011) and Gneiting and Ranjan (2013) to demonstrate a new method of assessing probabilistic calibration in empirical settings, whether the predictive CDFs are continuous or not, and with results that are not dependent on random draws from uniform distributions. We have also considered summary statistics that measure mis-calibration and produced one which can be decomposed into terms that measure mis-calibration primarily due to over/under-prediction and mis-calibration primarily due to over/under-dispersion.

The methods illustrated in this paper also apply to probability forecasts for binary events, since such forecasts can be represented with predictive CDFs. In the context of binary outcomes, the notion of conditional calibration (i.e., whenever the forecast probability of success is p, the event realizes with relative frequency p) is equivalent to that of probabilistic calibration (Gneiting and Ranjan, 2013). Conditional calibration has commonly been assessed using reliability diagrams (Murphy and Winkler, 1992), which is a plot of conditional observed frequencies against binned forecast probabilities. Any summary statistic that measures mis-calibration using this technique will depend on the choice of bins. In contrast, test statistics like $\mathrm{PS}_2(\bar{Q})$ that use the PIT eCDF $\bar{Q}$ provide a measure of conditional calibration independent of bin choice. It would be interesting to explore whether there are any direct connections between the PIT eCDF introduced in this report and the reliability curves generated from the statistically consistent optimal binning techniques of Dimitriadis et al. (2021).

In the classical setting, where predictive CDFs are continuous and PIT values are real numbers, formal tests of the hypothesis that a forecast system is probabilistically calibrated can be conducted. However, one must take into account any dependence structures in the set of PIT values. In the case of time series forecasts, one can sample sequential k-step ahead forecasts and use the autocorrelation function to test for independence within the subsequence prior to applying statistical tests for calibration (see Gneiting and Katzfuss

(2014) and the references therein). When dealing with discontinuous predictive CDFs, one could take a random draw from each CDF-valued PIT and apply classical methods on the draws. Such sampling and testing should be repeated many times to obtain an adequate distribution of test results. However, it would be preferable to devise analogues of these classical tests that apply directly to CDF-valued PITs. It is hoped that this paper encourages future research in this direction.

Acknowledgements

The author wishes to thank Ben Owen, Christopher Pickett-Heaps, Deryn Griffiths, Michael Foley and two anonymous reviewers who gave feedback on an earlier version of this manuscript.

A Proof of Proposition 1

Let $\bar{Q}$ denote the PIT eCDF for the sample. By symmetry,

$$\mathrm{PS}_2(\bar{Q}) = \int_0^1 (\bar{Q}^{-1}(x) - x)^2 \,\mathrm{d}x,$$

where $\bar{Q}^{-1}$ is the generalized inverse distribution function of $\bar{Q}$. Suppose that $U \sim V$, $W = \bar{Q}^{-1}(U)$, $Z = W - U$ and X is a random variable distributed by $\bar{Q}$ with X independent of U. Since $\mathrm{d}V(x) = \mathrm{d}x$,

$$\begin{aligned}
\mathrm{PS}_2(\bar{Q}) &= \mathbb{E}(Z^2) \\
&= \mathbb{E}\big((Z - \mathbb{E}(Z) + \mathbb{E}(Z))^2\big) \\
&= \mathbb{E}\big((Z - \mathbb{E}(Z))^2\big) + \mathbb{E}\big(2(Z - \mathbb{E}(Z))\mathbb{E}(Z)\big) + (\mathbb{E}(Z))^2 \\
&= \mathrm{Var}(Z) + (\mathbb{E}(Z))^2. \qquad \text{(A.1)}
\end{aligned}$$

Now

$$\begin{aligned}
\mathbb{E}(Z) &= \mathbb{E}(\bar{Q}^{-1}(U)) - \mathbb{E}(U) \\
&= \mathbb{E}(X) - \mathbb{E}(U), \qquad \text{(A.2)}
\end{aligned}$$

while

$$\begin{aligned}
\mathrm{Var}(Z) &= \mathrm{Var}(W) + \mathrm{Var}(U) - 2\,\mathrm{Cov}(W, U) \\
&= \mathrm{Var}(U) - \mathrm{Var}(W) + 2\big(\mathrm{Cov}(W, W) - \mathrm{Cov}(W, U)\big) \\
&= \mathrm{Var}(U) - \mathrm{Var}(X) + 2\,\mathrm{Cov}(W, W - U). \qquad \text{(A.3)}
\end{aligned}$$

Combining Equations (A.1), (A.2) and (A.3) completes the proof.

References

Anderson, J. L. (1996). A method for producing and evaluating probabilistic forecasts from ensemble model integrations. *J. Climate*, 9(7):1518–1530.

Bennett, J. C., Wang, Q. J., Robertson, D. E., Schepen, A., Li, M., and Michael, K. (2017). Assessment of an ensemble seasonal streamflow forecasting system for Australia. *Hydrology and Earth System Sciences*, 21(12):6007–6030.

Berthet, L., Bourgin, F., Perrin, C., Viatgé, J., Marty, R., and Piotte, O. (2020). A crash-testing framework for predictive uncertainty assessment when forecasting high flows in an extrapolation context. *Hydrology and Earth System Sciences*, 24(4):2017–2041.

Dawid, A. P. (1984). Present position and potential developments: Some personal views: statistical theory: the prequential approach. *Journal of the Royal Statistical Society: Series A (General)*, 147(2):278–290.

Diebold, F. X., Hahn, J., and Tay, A. S. (1998). Evaluating density forecasts with applications to financial risk management. *Int. Econ. Rev*, 39:863–883.

Dimitriadis, T., Gneiting, T., and Jordan, A. I. (2021). Stable reliability diagrams for probabilistic classifiers. *Proceedings of the National Academy of Sciences*, 118(8).

Gneiting, T., Balabdaoui, F., and Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):243–268.

Gneiting, T. and Katzfuss, M. (2014). Probabilistic forecasting. *Annual Review of Statistics and Its Application*, 1:125–151.

Gneiting, T. and Ranjan, R. (2013). Combining predictive distributions. *Electronic Journal of Statistics*, 7:1747–1782.

Hamill, T. M. (2001). Interpretation of rank histograms for verifying ensemble forecasts. *Mon. Wea. Rev.*, 129(3):550–560.

Hamill, T. M. and Colucci, S. J. (1997). Verification of Eta–RSM short-range ensemble forecasts. *Mon. Wea. Rev.*, 125(6):1312–1327.

Laio, F. and Tamea, S. (2007). Verification tools for probabilistic forecasts of continuous hydrological variables. *Hydrology and Earth System Sciences*, 11(4):1267–1277.

Li, W., Wang, Q. J., and Duan, Q. (2020). A variable-correlation model to characterize asymmetric dependence for postprocessing short-term precipitation forecasts. *Mon. Wea. Rev.*, 148(1):241–257.

Matheson, J. E. and Winkler, R. L. (1976). Scoring rules for continuous probability distributions. *Management science*, 22(10):1087–1096.

Murphy, A. H. and Winkler, R. L. (1987). A general framework for forecast verification. *Mon. Wea. Rev.*, 115(7):1330–1338.

Murphy, A. H. and Winkler, R. L. (1992). Diagnostic verification of probability forecasts. *International Journal of Forecasting*, 7(4):435–455.

Pearson, K. (1933). On a method of determining whether a sample of size n supposed to have been drawn from a parent population having a known probability integral has probably been drawn at random. *Biometrika*, pages 379–410.

Raftery, A. E., Gneiting, T., Balabdaoui, F., and Polakowski, M. (2005). Using Bayesian model averaging to calibrate forecast ensembles. *Mon. Wea. Rev.*, 133(5):1155–1174.

Renard, B., Kavetski, D., Kuczera, G., Thyer, M., and Franks, S. W. (2010). Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors. *Water Resources Research*, 46(5).

Rosenblatt, M. (1952). Remarks on a multivariate transformation. *The annals of mathematical statistics*, 23(3):470–472. https://www.jstor.org/stable/2236692.

Shao, Y., Wang, Q. J., Schepen, A., and Ryu, D. (2021). Going with the trend: forecasting seasonal climate conditions under climate change. *Mon. Wea. Rev.*, 149(8):2513–2522.

Talagrand, O. (1999). Evaluation of probabilistic prediction systems. In *Proceedings of Workshop on Predictability*. 20-22 October 1997, ECMWF, Reading, UK.

Thyer, M., Renard, B., Kavetski, D., Kuczera, G., Franks, S. W., and Srikanthan, S. (2009). Critical evaluation of parameter consistency and predictive uncertainty in hydrological modeling: A case study using bayesian total error analysis. *Water Resources Research*, 45(12).

Wang, Q. and Robertson, D. (2011). Multisite probabilistic forecasting of seasonal flows for streams with zero value occurrences. *Water Resources Research*, 47(2).

Wang, Q., Robertson, D., and Chiew, F. (2009). A bayesian joint probability modeling approach for seasonal forecasting of streamflows at multiple sites. *Water Resources Research*, 45(5).

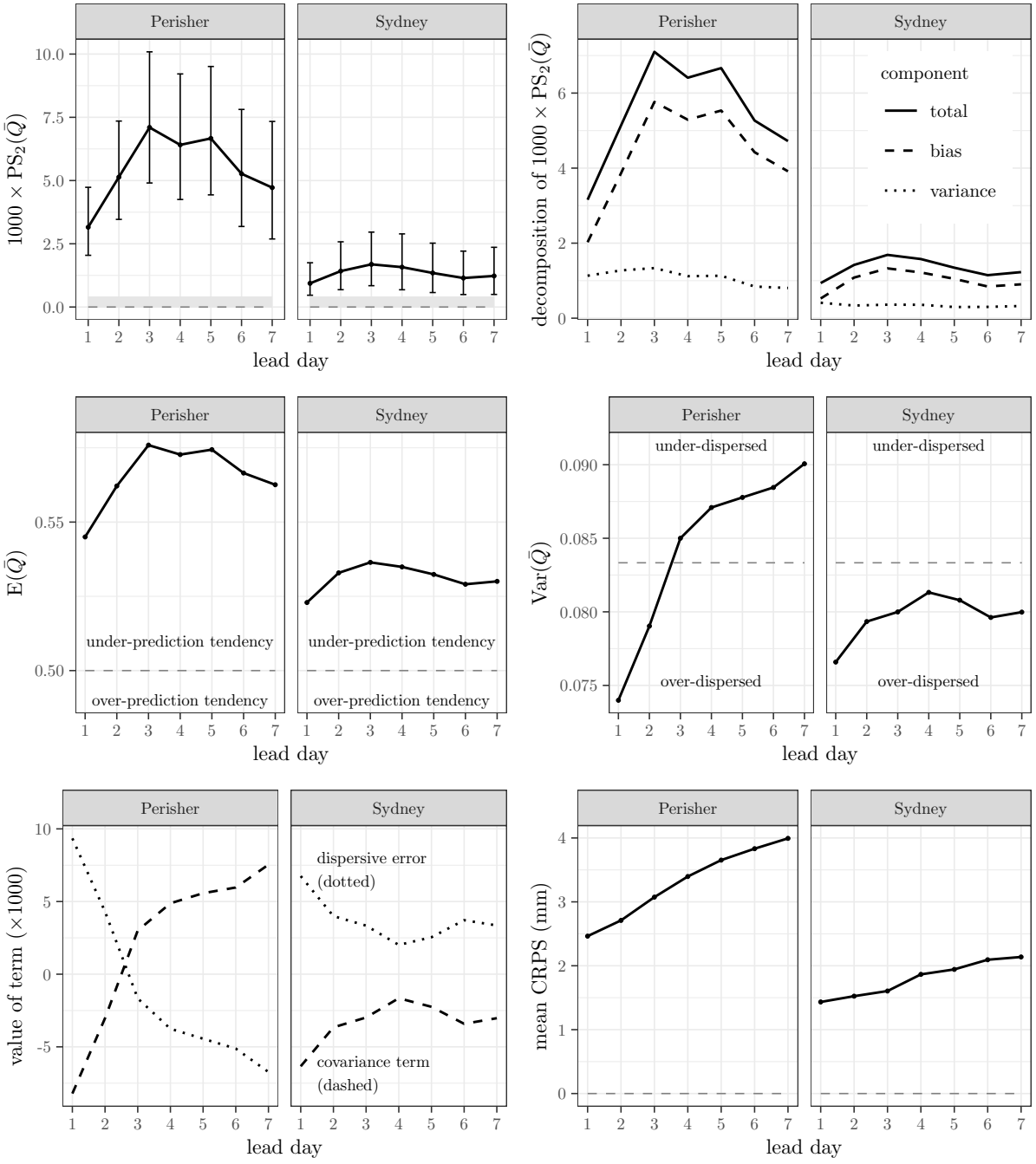

Figure 8: Calibration and accuracy by forecast lead day at Perisher Valley and Sydney Airport for daily precipitation forecasts issued by the BoM. Top left: $\mathrm{PS}_2(\bar{Q})$ with bootstrapped 95% intervals. The upper boundary of the shaded gray region is the critical value for the Cramér–von Mises significance test for uniformity at the 5% level. Top right: Decomposition of $\mathrm{PS}_2(\bar{Q})$ (solid lines) into bias $(\mathbb{E}(Y) - \mathbb{E}(U))^2$ (dashed lines) and variance $\mathrm{Var}(\bar{Q}^{-1}(U) - U)$ (dotted lines) components. Centre left: Bias, measured by $\mathbb{E}(\bar{Q})$ relative to 0.5. Centre right: Dispersion, measured by $\mathrm{Var}(\bar{Q})$ relative to 1/12. Bottom left: Values of dispersive error (dotted lines) and covariance (dashed lines) terms in the second decomposition of $\mathrm{PS}_2(\bar{Q})$. Bottom right: mean CRPS.