Australian Government
Bureau of Meteorology

Ensemble transformations for consistency with a target forecast and its application to seamless weather to subseasonal forecasting

Robert J. Taggart, Morwenna Griffiths, Matthew C. Wheeler, Claire M. Spillman

January 2025

Bureau Research Report - 104

Ensemble transformations for consistency with a target forecast and its application to seamless weather to subseasonal forecasting

Robert J. Taggart, Morwenna Griffiths, Matthew C. Wheeler, Claire M. Spillman

Bureau of Meteorology

Bureau Research Report No. 104

January 2025

National Library of Australia Cataloguing-in-Publication entry

Authors: Robert J. Taggart, Morwenna Griffiths, Matthew C. Wheeler, Claire M. Spillman

Title: Ensemble transformations for consistency with a target forecast and its application to seamless weather to subseasonal forecasting

ISBN: 978-1-925738-90-2

ISSN: 2206-3366

Series: Bureau Research Report – BRR104

Enquiries should be addressed to:

Lead Author: Robert J. Taggart

Bureau of Meteorology

GPO Box 1289, Melbourne

Victoria 3001, Australia

robert.taggart@bom.gov.au

Copyright and Disclaimer

Contents

List of Figures

List of Tables

Executive summary

It is common for forecasts like accumulated precipitation or average temperature over a period to be produced by different forecast systems, depending on the forecast lead time and period duration. For example, 24-hour precipitation accumulations out to 7 days are typically produced by a different system than weekly precipitation accumulations out to 4 weeks. To provide a seamless forecast service over all periods and lead times, outputs from one system need to be post-processed to be consistent with the other system for the overlapping prediction period, or outputs from both need to be blended together to form a new forecast. This report presents methods for post-processing forecasts from an ensemble prediction system (EPS) so that its forecasts are consistent with statistically well-defined forecasts produced by a second forecast system, called the target system. We provide methods for two cases: (1) when the cumulative distribution function of the target system is known or can be reconstructed, and (2) when the target system only provides a mean (i.e., expected) value forecast. These methods are illustrated for precipitation and air temperature using two forecast systems from the Australian Bureau of Meteorology. It is found that enforcing consistency of daily EPS forecasts at lead days 1 to 7 results in week 1 (e.g., 7-day precipitation accumulations) forecasts being substantially more accurate than week 1 forecasts from either the original EPS or the target forecast system.

One of the methods for enforcing consistency requires the values from each ensemble forecast to be fit to a parametric distribution. Computationally efficient methods are presented for generating parametric fits to EPS forecasts of precipitation and temperature over large gridded datasets. Parametric fitting has many uses, so these methods are likely to be of interest to other environmental forecast applications.

1 Introduction

Many users of environmental forecasts make decisions that are dependent on forecast information that span a range of lead times. For example, agricultural users frequently consider both short-term weather and subseasonal-to-seasonal climate predictions for their decisions (Mase and Prokopy, 2014). The term "seamless prediction" is used to describe the integration of weather–climate forecasts across multiple time scales in the range of minutes to years (Shukla, 2009; Hoskins, 2013; Brunet et al., 2015; Ren et al., 2023). The ideal is to provide users forecast information that shows no boundaries between different time scales due to the use of different techniques and/or prediction systems (Kumar and Murtugudde, 2013; Ruti et al., 2020). One approach to tackling this problem is a seamless modelling approach whereby the same model or model family is used across time scales (Senior et al., 2008; Brown et al., 2012). For example, in 2008 the European Centre for Medium Range Weather Forecasts chose to run the same atmosphere model for both medium-range and monthly forecasting, but with reduced model resolution and coupled to an ocean from day 10 (Vitart et al., 2008).

Despite this progress, different forecast systems are still often used to generate forecasts for different prediction horizons. In this report we therefore present a statistical approach to making seamless forecast information across these different systems, providing methods that can be used to transform forecasts from an ensemble prediction system (EPS) so that they are consistent with forecasts from a target system in the intersecting prediction period. Essentially, our methods can be applied when forecasts from the target system can be expressed as cumulative distribution functions (CDFs) or are mean (i.e. expected) values of some underlying predictive distribution. The methods transform forecasts from the EPS to have distributional properties that are consistent with the target distribution. Forecasts from the target system are not adjusted, unlike other statistical approaches that *blend* forecasts by taking a weighted mean of forecasts from two systems to generate a new set of forecasts with seamless properties (Kober et al., 2012; Scheufele et al., 2014). A forecast system might be considered a target because it has high usage yet one does not have permission to modify its predictions, or because it is has very high predictive skill relative to the EPS where their forecast horizons overlap.

To illustrate the need for seamless prediction and why such methods have been developed, we introduce two major forecast systems from the Australian Bureau of Meteorology (henceforth 'the Bureau') covering the Australian domain: the Australian Digital Forecast Database (ADFD) and Australian Community Climate and Earth System Simulator – Seasonal (ACCESS-S). The ADFD publishes detailed weather forecasts from days 0 to 7 while ACCESS-S is primarily used to generate subseasonal to seasonal forecasts. Predictions from both forecast systems are published on the Bureau's website (`http://www.bom.gov.au/`) and overlap in the week 1 prediction horizon. Moreover, both systems have different production methods and outputs. The ADFD is based on a statistically post-processed blend of numerical weather prediction (NWP) models (Bureau of Meteorology, 2018; Griffiths and Jayawardena, 2022; Trotta et al., 2024), manually curated by operational meteorologists (Just and Foley, 2020). The ADFD outputs several probability of exceedance (PoE) and percentile forecasts for daily precipitation, and mean-value forecasts for daily precipitation, maximum screen temperature (Tmax) and minimum screen temperature (Tmin) (Bureau of Meteorology, 2023). On the other hand,

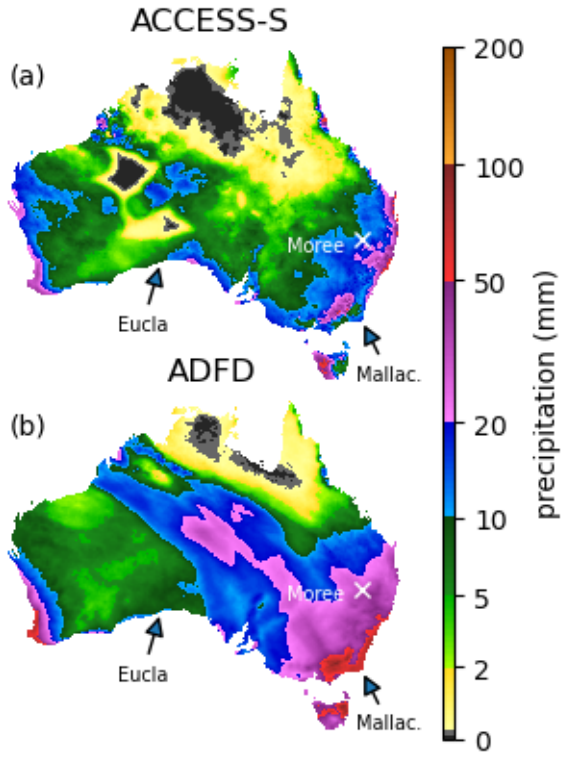

Figure 1: Mean value 7-day precipitation accumulation for the period 30 May to 5 June 2024 forecast by (a) the ACCESS-S ensemble and (b) ADFD. Arrows point to the coastal localities of Mallacoota and Eucla, while the cross indicates the inland location of Moree Airport.

ACCESS-S is a partially calibrated 99-member EPS produced by a coupled dynamical atmosphere–ocean model (Hudson et al., 2017; Wedd et al., 2022; Griffiths et al., 2023). While one cannot expect forecasts from these two different forecast systems to precisely agree, at times ADFD and ACCESS-S forecasts present obviously inconsistent stories in the overlapping week 1 period, as has occasionally been noted in the media (Saunders, 2023). For example, Fig. 1 shows mean value forecasts of 7-day accumulated precipitation issued from ACCESS-S and ADFD on 29 May 2024 for the 7-day period 30 May to 5 June 2024. The ADFD forecast is wetter than ACCESS-S across most of Australia, and at Mallacoota (location indicated in Fig. 1) forecasts a value of 63 mm compared with 16 mm from ACCESS-S. The ADFD is a classic candidate for a target forecast since current generation processes, particularly manual input from operational meteorologists, rule out automated blending with ACCESS-S as a final step.

This report gives seamless transformations of EPSs for two particular cases.

- Case 1: transform the values from an EPS so that they are consistent with a given target predictive CDF (Section 2).
- Case 2: transform the values from an EPS so that the ensemble mean equals a target mean value forecast (Section 3).

Two methods address Case 1. The first method essentially forces the ensemble to match the target CDF, with the effect that the transformed ensemble members are evenly sampled (in probability space) from the target distribution. We call this type of transformation *strongly consistent*. The second method fits the ensemble to a parametric distribution and then uses quantile–quantile mapping between the parametric and target CDFs to transform the ensemble. The transformed ensemble will typically look like a random sample from the target distribution. We call this type of transformation *weakly consistent*. If probabilistic consistency with the target distribution is the primary goal, then the first method is preferred. However, if preserving some of the temporal and cross-parameter dependencies within the ensemble is desired, the second method should be considered.

Of particular interest is how the methods handle the situation where all members of the ensemble forecast dry conditions (0 mm) but the target forecast has rainfall. The first method handles this using a ranking function which determines which transformed ensemble members will be assigned rainfall. The choice of ranking function has no impact on consistency with the target forecast but does impact spatial properties of transformed ensemble members and predictive accuracy for precipitation accumulated over multiple time steps. The second method can be slightly altered to handle this situation using the same ranking technique. These issues are explored in Section 5.3.

Weakly consistent transformations require fitting ensemble data to suitable parametrized distributions. Operational implementation of this method requires efficient estimation of these parameters for large gridded datasets. For temperature forecasts, the normal, skew normal, mixture of two normals, and beta distributions can collectively offer good fits. For precipitation forecasts, the gamma distribution offers good fits for nonzero data. Appendices A and B present methods for parameter estimation for these families of distributions that can be vectorized (i.e., can compute parameters for many grid cells simultaneously in parallel) and are thus operationally efficient. They are based on the method of moments, other methods from the statistics literature (Balakrishnan and Wang, 2000; Arnold et al., 1993; Tan and Chang, 1972) or modifications thereof. Curve fitting algorithms for ensemble forecasts are used for other applications (National Weather Service, 2024; de Burgh-Day and Dillon, 2021) and so the efficient techniques presented in the appendix are likely to be of wider interest.

The situation of Case 2 could be addressed by translating the entire ensemble so that the ensemble mean equals the target forecast. However, this can result in implausible transformed ensemble member values, such as negative values for precipitation or values that lie well outside the climatic range for a location. Instead, we introduce a transformation of the ensemble that respects pre-specified bounds, ensures that the transformed ensemble mean equals the target forecast, and where possible preserves ensemble variance.

Both Cases 1 and 2 and their corresponding methods are illustrated using forecasts from the ADFD (the target forecast) and the ACCESS-S ensemble. Details of the forecast and observation datasets, as well as verification methods, are given in Section 4. Case 1 is then illustrated for daily precipitation for days 1 to 7 using CDFs constructed from ADFD probabilistic information as the target distribution (Section 5). Case 2 is illustrated using daily Tmin and Tmax for days 1 to 7, using the ADFD mean value as the target (Section 6). When assessed for predictive accuracy, the ADFD is generally better at short range daily prediction, but has the disadvantage that forecasts for multi-day accumulations (for precipitation) or multi-day means (for temperature) are single-valued

and not in the form of a CDF. Using the seamless transformation methods of this report, the transformed ensemble can be used to generate CDFs for multi-day prediction. These seamless multi-day CDFs have substantially superior predictive skill than either the original ACCESS-S ensemble or the ADFD single-valued alternative. As such, the authors have implemented an additional post-processing step within the research production pipeline that transforms ACCESS-S precipitation forecasts so that they are strongly consistent with ADFD precipitation forecasts. A similar post-processing step is in preparation for ACCESS-S temperature forecasts. Once implemented operationally, the improved predictive skill of week 1 forecasts from ACCESS-S, plus the increased consistency between the Bureau's forecasts from weather to subseasonal time scales, is likely to enhance user trust in the Bureau's predictive services (Burgeno and Joslyn, 2020).

2 Case 1: Adjusting an EPS for consistency with a target predictive distribution

2.1 Notions of consistency

Suppose that Y is some unknown quantity whose potential values come from (a subset of) the real numbers $\mathbb{R}$. For example, Y could be the accumulated precipitation or daily Tmax for a particular location and date. A predictive distribution for Y is a CDF F which, for any given threshold $\theta \in \mathbb{R}$, gives the probability of non-exceedance for Y, namely, $F(\theta) = \mathbb{P}(Y \leq \theta)$. An n-member EPS forecast for Y is a vector $\mathbf{x}$ in $\mathbb{R}^n$, given by $\mathbf{x} = (x_i)_{i=1}^n$, where x_i is the prediction from the ith member of the ensemble. Denote by $(x_{(j)})_{j=1}^n$ the reordered ensemble satisfying $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$, so that $x_{(j)}$ is the jth smallest value from among the ensemble members.

Suppose that we have two different predictions for Y, one in the form of a CDF F and the other in the form of a vector $\mathbf{x}$ from an EPS. Fig. 2a illustrates the graph of a CDF F for daily Tmax (the gold curved line) and corresponding forecasts $\mathbf{x}$ from an EPS (the light blue crosses) with ensemble size $n = 33$, with each ensemble member $x_{(j)}$ plotted using the plotting position $(x_{(j)}, j/(n+1))$. If $\mathbf{x}$ is interpreted as a sample from some underlying distribution G, it seems implausible that F is the same as, or even close to, G. For example, the median value of F is 35.3°C while the median of $\mathbf{x}$ is 26.2°C. We can say that the forecasts F and $\mathbf{x}$ are *inconsistent*. We now make notions of consistency and inconsistency a little more explicit.

A forecast $\mathbf{x}$ from an EPS is said to be

1. *strongly consistent* with a predictive distribution F if $\lim_{z \uparrow x_{(j)}} F(z) \leq j/(n+1) \leq F(x_{(j)})$;

2. *weakly consistent* with a predictive distribution F if the values of $\mathbf{x}$ are a plausible sample from the distribution F.

Strong consistency implies that ordered ensemble $(x_{(j)})_{j=1}^n$ consists of quantiles whose levels are evenly spaced on the distribution F, with the left-sided limit in the definition required to handle cases where F is discontinuous. To make weak consistency well-defined, the notion of 'plausible sample' requires further elaboration. Plausibility could be said to occur if the null hypothesis that 'the sample comes from the distribution' is not rejected by

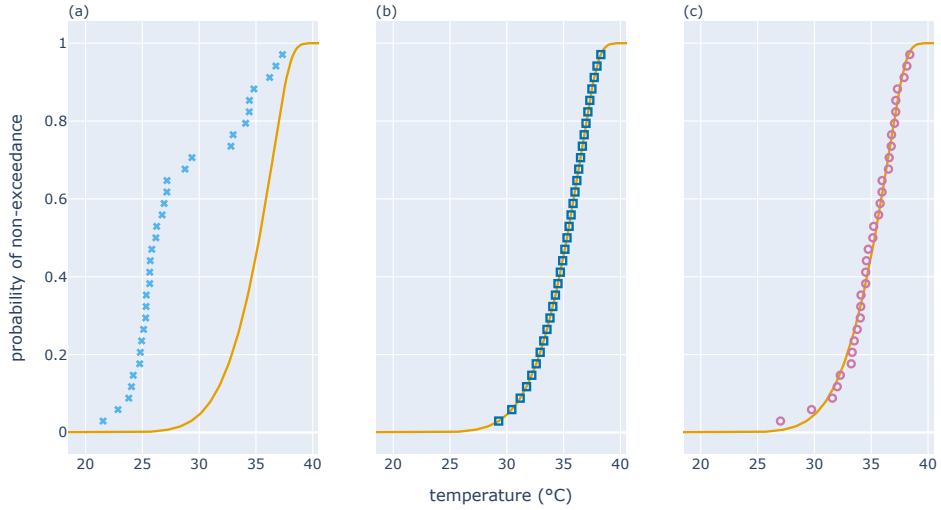

Figure 2: Three different ensemble forecasts (illustrated with markers) of size $n = 33$ which are respectively (a) inconsistent, (b) strongly consistent and (c) weakly consistent with the CDF of the target distribution F (gold curved line).

(say) the Kolmogorov–Smirnov test at the 5% significance level. However, for this report a precise definition of weak consistency will not be required. Instead, we focus on whether the method that invokes weak consistency produces better forecasts for the application at hand. Strong consistency implies weak consistency, and an ensemble forecast $\mathbf{x}$ that is not weakly consistent is said to be *inconsistent* with F.

Fig. 2 illustrates three different ensemble forecasts $\mathbf{x}$ of Tmax of size $n = 33$ and compares them with a predictive distribution F whose CDF is depicted by a gold curve. It has already been noted that the first EPS forecast (light blue crosses in Fig. 2a) is inconsistent with F. If we define an EPS forecast to be weakly consistent using the Kolmogorov–Smirnov test at the 5% significance level, then the null hypothesis is rejected with a p-value of 1.6×10^{-15}. Fig. 2b illustrates an EPS forecast (dark blue squares) that is strongly consistent with F: the squares lie on the graph of F at evenly spaced quantile levels. The EPS forecast (pink circles) in Fig. 2c shows some slight scatter about the gold line and is illustrative of weak consistency. Indeed, the null hypotheses that the samples $\mathbf{x}$ in Figs. 2b and 2c come from the distribution F are accepted at the 5% significance level with p-values of 1.0 and 0.96 respectively.

2.2 Two methods to generate consistent ensembles

This report presents two methods to transform a vector $\mathbf{x}$ of ensemble forecasts so that it is consistent with a target predictive distribution F. The first method transforms $\mathbf{x}$ so that it is strongly consistent with F, while the second (typically) results in a transformed ensemble that is weakly consistent with F. Both methods involve quantile–quantile matching, but

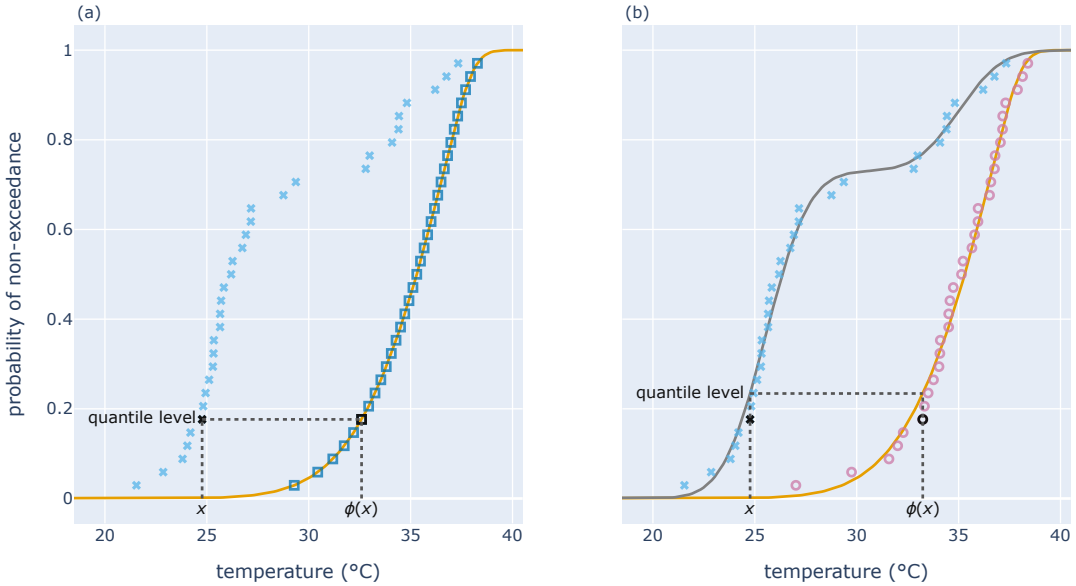

Figure 3: Illustration of two transformation techniques, applied to a 33-member ensemble (blue crosses) that is inconsistent with the target distribution F (gold curve). In (a) the ensemble is transformed to be strongly consistent with F (dark blue squares), while in (b) it is transformed to be weakly consistent with F (pink circles) using a fitted distribution (gray curve). In both panels, x is one value from the inconsistent ensemble and $\phi(x)$ is the transformed value in the consistent ensemble.

differ in how each quantile level is estimated. Both are illustrated in Fig. 3, using the inconsistent EPS forecast $\mathbf{x}$ from Fig. 2a as a starting point and, post-transformation, producing either the strongly or weakly consistent forecasts of Figs. 2b and 2c.

The first method requires a ranking function R that assigns a unique rank (that is, a number from 1 to n) to each of the ith components from an n-member ensemble $\mathbf{x}$. Specifically, we write $R(\mathcal{I}, i) = j$ to indicate that R assigns the ith ensemble member the rank j using the information set $\mathcal{I}$.

For example, if R assigns rank based on the original ensemble values, where a higher value is assigned a higher rank, then the information set $\mathcal{I}$ is the ensemble forecast $\mathbf{x}$. Suppose that $\mathbf{x} = (5.1, 9.7, 0.3)^T$. Then $R(\mathcal{I}, 2) = 3$ since x_2 is the highest value, while $R(\mathcal{I}, 3) = 1$ since x_3 is the smallest value. Ranking functions R that respect the ordering of the initial ensemble forecast $\mathbf{x}$ have the property that $\mathbf{x}$ belongs to the information set $\mathcal{I}$ and that $R(\mathcal{I}, i) < R(\mathcal{I}, j)$ whenever $x_i < x_j$. However, $\mathbf{x}$ may have tied values, in which case ties could be broken using additional information or by random methods. For example, to rank two daily precipitation values of 0 mm in $\mathbf{x}$, R could use cloud cover or humidity forecasts from the EPS to break the tie.

Given a ranking function R, we define a mapping ϕ_s that takes a target CDF F and

information set $\mathcal{I}$ and returns an ensemble forecast $\phi_{\mathrm{s}}(\mathcal{I}, F; R)$ that is strongly consistent with F via the formula

$$\phi_{\mathrm{s}}(\mathcal{I}, F; R) = \left(F^{-1}\left(\frac{R(\mathcal{I}, i)}{(n+1)} \right) \right)_{i=1}^{n}, \tag{1}$$

where F^{-1} is the quantile function (i.e., generalized inverse distribution function) of F. In other words, if $\tilde{\mathbf{x}}$ denotes the ensemble forecast given by $\tilde{\mathbf{x}} = \phi_{\mathrm{s}}(\mathcal{I}, F; R)$, then jth largest member $\tilde{x}_{(j)}$ of $\tilde{\mathbf{x}}$ is the $j/(n+1)$-quantile value of F. Hence $\tilde{\mathbf{x}}$ is strongly consistent with F. In the case where the information set is precisely the original ensemble forecast $\mathbf{x}$, we have the strongly consistent transformation $\tilde{\mathbf{x}} = \phi_{\mathrm{s}}(\mathbf{x}, F; R)$.

Fig. 3a illustrates how the quantile–quantile matching works for ϕ_{s} when R ranks ensemble members based on the ordering of their values. The 6th ranked ensemble member (shown with an x in Fig. 3a) has a value of 24.8 is assigned a quantile level of 6/34. It is then mapped to the (6/34)-quantile value of F which is 32.6 (shown as $\phi(x)$ in Fig. 3a).

The choice of ranking function R can effect (1) spatial characteristics of transformed ensemble members and (2) predictive accuracy of aggregated forecasts (e.g. accumulated rainfall or average temperature over multiple time steps) from the transformed ensemble. This is explored further when ϕ_{s} is applied to Bureau precipitation forecasts in Section 5.3.

The second method requires an algorithm $\mathcal{F}$ to 'fit' the ensemble forecast $\mathbf{x}$ to a parametric distribution. That is, $\mathcal{F}$ takes a sample $\mathbf{x}$ of n data points and outputs a CDF $\mathcal{F}(\mathbf{x})$ that comes from one or more families of parametric distributions. Given a fitting algorithm $\mathcal{F}$, we define a transformation ϕ_{w} that takes an EPS forecast $\mathbf{x}$ and a target CDF F and returns a transformed forecast $\phi_{\mathrm{w}}(\mathbf{x}, F; \mathcal{F})$ by the formula

$$\phi_{\mathrm{w}}(\mathbf{x}, F; \mathcal{F}) = \left(F^{-1}\left((\mathcal{F}(\mathbf{x}))(x_i) \right) \right)_{i=1}^{n}. \tag{2}$$

The transformation ϕ_{w} works as follows. First the ensemble forecast $\mathbf{x}$ is fitted to a distribution whose CDF G is given by $G = \mathcal{F}(\mathbf{x})$. Given an ensemble member x_i, the quantile level α_i of x_i is calculated by $\alpha_i = G(x_i)$. The transformed value of x_i is the α_i-quantile of the target distribution F, namely $F^{-1}(\alpha_i)$. If G is a reasonable fit to the original ensemble, then the resulting transformed ensemble will appear a plausible sample of the target distribution and hence the transformation will be weakly consistent.

Fig. 3b illustrates how the quantile–quantile matching works for ϕ_{w}. The original ensemble forecast (blue crosses) is fitted to a distribution G whose graph is shown (grey curve). For the 6th ranked ensemble member ($x = 24.8$), the estimated quantile level is $G(x) = 0.22$ and the transformed ensemble member is $\phi(x) = F^{-1}(0.22) = 33.3$.

The weakly consistent transformation ϕ_{w} preserves the anomalous scatter of the ensemble about the fitted distribution G when it remaps it about the distribution F. We can make this precise by defining the quantile level anomaly (QLA) of an ensemble member $x_{(j)}$ with respect to G to be $G(x) - j/(n+1)$. That is, the QLA of x with respect to G is the difference between the quantile level $G(x)$ and an unbiased estimator of the quantile level based on rank. It can be visualized in Fig. 3 as the signed vertical distance between the plotted ensemble member and graph of the corresponding CDF. The QLAs of the original ensemble relative to the fitted distribution are the same as the QLAs for the ϕ_{w}-transformed ensemble relative to the target distribution. Visually, if a blue cross lies above (respectively below) the gray curve, then the corresponding pink circle will also lie above (respectively below) the gold curve, and thus ϕ_{w} preserves more information from

the original ensemble in its transformation than ϕ_s. In contrast, ϕ_s 'snaps' the ensemble to the target distribution and the natural variability within the original ensemble is lost.

From a practical perspective, the fitting algorithm $\mathcal{F}$ used in ϕ_w should (1) be computationally efficient at estimating distribution parameters over large gridded datasets, to meet operational time constraints within a production setting, (2) typically produce good fits to the ensemble data so that the adjusted ensemble is weakly consistent with the target forecast, and (3) not have too many parameters in the family of distributions to avoid over-fitting of data with resulting unrealistic appraisals of QLAs. Fitting algorithms that are suitable for use with precipitation and temperature forecasts are discussed in Appendices A and B.

Some meteorological forecast variables, such as precipitation and wind magnitude, have distributions that are best modelled by a mix of discrete and continuous distributions. For precipitation, the discrete component is the likelihood that it rains while the continuous component is the distribution of precipitation given that rain occurs. It is possible to modify the weakly consistent transformation so that the transformed ensemble is strongly consistent with the discrete component of the target distribution, and weakly consistent with the continuous component. This is illustrated in Section 5 using Bureau precipitation forecasts.

3 Case 2: Adjusting an EPS for consistency with a target mean value forecast

We turn now to the situation where the target forecast is the mean (i.e., expected) value z from some predictive distribution, but the predictive distribution itself is not known. The aim is to transform the ensemble $\mathbf{x}$ so that it is consistent with z. In this case, consistency means that the mean of the transformed ensemble should equal z.

An initial approach is to translate the ensemble $\mathbf{x}$ using the transformation ϕ_t, where

$$\phi_t(\mathbf{x}, z) = \mathbf{x} - \bar{\mathbf{x}} + z$$

and $\bar{\mathbf{x}}$ denotes the mean value of $\mathbf{x}$, so that the transformed ensemble mean $\overline{\phi_t(\mathbf{x}, z)}$ equals the target z. However, ϕ_t could map some ensemble members to unrealistic values, falling well outside the climatological range for particular location, or outside the allowable physical range of a parameter (such as values for wind magnitude less than 0 ms^{-1}).

To avoid this situation, we introduce a new transformation ϕ_b that behaves like ϕ_t much of the time but ensures that the output respects bounds as specified by a tuple $B = (\ell, u, b_1, b_2)$ of four parameters. The lower bound ℓ and upper bound u parameters specify the range $[\ell, u]$ of acceptable values for the variable being forecast. The non-negative buffer parameters b_1 and b_2 are used to expand the range of acceptable values should z fall outside $[\ell, u]$ or come within a distance b_1 of ℓ or a distance b_2 of u. Given the bound parameters B, define the transformation ϕ_b by

$$\phi_b(\mathbf{x}, z; B) = c(\mathbf{x} - \bar{\mathbf{x}}) + z, \quad \text{where} \quad c = \min\left(1, \frac{z - L(z)}{\bar{\mathbf{x}} - x_{(1)}}, \frac{U(z) - z}{x_{(n)} - \bar{\mathbf{x}}}\right), \tag{3}$$

$L(z) = \min(\ell, z - b_1)$ and $U(z) = \max(u, z + b_2)$. In the case when $x_i = \bar{\mathbf{x}}$ for every i, we interpret c to take the value 1.

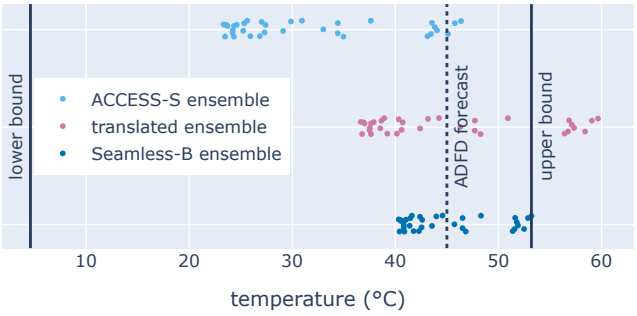

Figure 4: Lead day 5 daily Tmax forecasts for 23 February 2023 at Eucla. First row: forecasts from the 33-member ACCESS-S ensemble. Second row: translation of ACCESS-S using ϕ_t so that the ensemble mean matches the ADFD forecast of 45.0°C. Third row: Seamless-B forecast, which is a transformation of ACCESS-S using ϕ_b so that the ensemble mean matches the ADFD and respects pre-specified bounds.

If $\mathbf{v} = \phi_b(\mathbf{x}, z; B)$ and z is the target forecast then the transformation ϕ_b and the transformed ensemble $\mathbf{v}$ have the following properties:[1]

1. The transformation ϕ_b is order preserving in the sense that $v_i \leq v_j$ whenever $x_i \leq x_j$.
2. The mean value $\bar{\mathbf{v}}$ of the transformed ensemble is z.
3. The variance $\text{var}(\mathbf{v})$ of the transformed ensemble is $c^2\,\text{var}(\mathbf{x})$.
4. The transformed ensemble $\mathbf{v}$ always lies within the range $[L(z), U(z)]$; that is, $L(z) \leq v_i \leq U(z)$ for all i.
5. If all components of the translated ensemble $\phi_t(\mathbf{x}, z)$ lie in the interval $[L(z), U(z)]$, then $c = 1$ and $\mathbf{v} = \phi_t(\mathbf{x}, z)$.

In summary, ϕ_b can be used to transform an ensemble so that its mean matches the desired target and each ensemble member lies in the range $[L(z), U(z)]$, and where possible ϕ_b produces the same outputs as the translation operator ϕ_t.

We illustrate the transformations ϕ_t and ϕ_b using the 33-member[2] ACCESS-S ensemble Tmax forecast for 23 February 2023 at Eucla (location shown in Fig. 1), with the ADFD mean as the target forecast. The ACCESS-S forecast is shown with light blue markers in Fig. 4, with ensemble members ranging from 23.3°C to 46.4°C (light blue markers), depending on whether forecast winds are onshore (resulting in lower temperatures) or

[1]To prove property Property 1 it suffices to show that $c \geq 0$. Property 2 follows from the linearity of the expectation operator. Property 3 follows from basic properties of the variance operator. The proofs of Properties 4 and 5 use the fact that $c < 1$ if and only if $x_{(1)} - \bar{\mathbf{x}} + z < L(z)$ or $x_{(n)} - \bar{\mathbf{x}} + z > U(z)$.

[2]The 33-member ensemble, rather than the 99-member time lagged ensemble, is used for illustrative purposes so that the diagram is less cluttered.

offshore (leading to higher temperatures). The ADFD mean value forecast z is 45.0°C (dotted vertical line). When the original ensemble is translated using ϕ_t (pink markers) so that the ensemble mean matches the ADFD forecast, seven ensemble members exceed 55°C and the highest ensemble member is 59.7°C. These forecasts are not plausible for Eucla's climate. Instead, we use the transformation ϕ_b with bound parameters $B = (\ell, u, b_1, b_2) = (4.6, 53.2, 1, 1)$ (in °C) to create the 'Seamless-B' ensemble forecast (dark blue markers). The chosen lower and upper bounds ℓ and u are based on climatology (see Section 6.1 for details). While both the translated and Seamless-B ensemble means equal the ADFD target forecast, only the Seamless-B ensemble forecast respects the upper bound. It achieves this by reducing the ensemble range to 40.3°C to 53.2°C.

Note that the transformation ϕ_b is designed to generate ensemble forecasts with certain statistical properties. There is no guarantee that the transformed variable (in this case Tmax) will retain dynamical or meteorological consistency with other parameters (e.g. wind direction) within each ensemble member.

4 Datasets and verification methods

The methods for adjusting an EPS for consistency with a target forecast will be applied, in Sections 5 and 6, to precipitation, Tmax and Tmin forecasts from the ACCESS-S ensemble with target forecasts provided by the ADFD. In this section we give details of the observation and forecast datasets and of the verification method used to assess predictive accuracy. The assessment validity periods are 23 November 2023 to 13 November 2024 for daily precipitation and 11 November 2022 to 10 November 2023 for daily Tmin and Tmax.

4.1 Observations

Observations were obtained from the Bureau's network of automatic weather stations (AWSs) in Australia, with a resolution of 0.1°C for temperature and 0.2 mm for precipitation. The locations of the AWSs used are shown in Fig. 5, with 522 providing daily precipitation observations and 502 providing temperature observations.

Observational validity periods match the validity periods of official Bureau (and hence ADFD) day 1 to 7 forecasts across Australia as follows. The observed daily precipitation at a location is the 24-hour accumulation starting at 15 UTC (approximately midnight local time). The observed daily Tmin at a location is the minimum observed temperature for the 18-hour period from 07 UTC (early evening local time) to 01 UTC (late morning local time). The observed daily Tmax at a location is the maximum observed temperature for the 18-hour period from 19 UTC (early morning local time) to 13 UTC (late evening local time). Observations are quality controlled using automated processes from the Bureau's Jive forecast verification system (Loveday et al., 2024). For each parameter, fewer than 5% of observations were missing or removed by quality control processes.

4.2 ACCESS-S forecasts

ACCESS-S uses a coupled atmosphere–ocean model on an approximately 60km horizontal grid for the atmosphere (Wedd et al., 2022; Hudson et al., 2017). For time horizons out

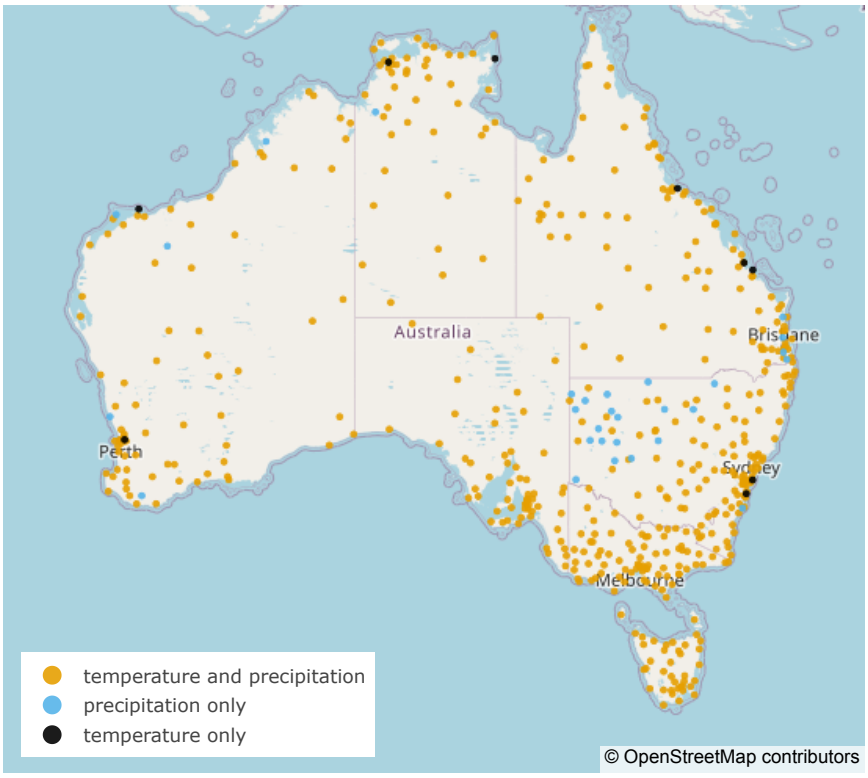

Figure 5: Location of AWSs used for the evaluation of precipitation and temperature forecasts in this study.

to six weeks, 33 ensemble members are run daily and from those a 99-member time-lagged ensemble is constructed. For a subset of atmospheric variables, each ensemble member is downscaled to a 5km grid over Australia (land only) and calibrated using quantile–quantile matching as described by Griffiths et al. (2023). Downscaled, calibrated 99-member ensemble forecasts for daily precipitation and temperature are used in this study, with values corresponding to AWS observations selected using the nearest grid cell to the AWS location.

The validity period for 24-hour daily precipitation forecasts from ACCESS-S start at 15 UTC, matching validity periods for the observations and ADFD. The validity periods for ACCESS-S Tmax and Tmin forecasts both start at 00 UTC and have a duration of 24 hours. Although this differs from the validity periods of ADFD Tmax and Tmin forecasts and corresponding observations, this was not seen as a significant drawback, since the time that the daily maximum or minimum is attained typically falls in the intersection of the ADFD and ACCESS-S validity periods.

Data archival issues resulted in a little over 80% availability of the 99-member ACCESS-S forecasts for days 1 to 7 for the verification periods in this study.

4.3 ADFD forecasts

Daily temperature and precipitation forecasts published by the Bureau to the ADFD cover the Australian domain on an approximately 6km spatial resolution to lead day 7. The

Table 1: Daily precipitation forecasts published to the ADFD.

Type of forecast	details
Quantile forecast at the α level	$\alpha \in \{0.25, 0.5, 0.75, 0.9\}$
Probability of exceeding θ mm	$\theta \in \{0, 1, 5, 10, 15, 25, 50\}$
Mean (i.e., expected) value	

elevation of any ADFD grid cell that contains an AWS is set to be the same elevation as the AWS, and ADFD forecasts at these grid cells were sampled for comparison with AWS observations.

To present and compare spatial outputs (e.g. Fig. 1), ADFD forecasts were regridded to the same 5 km grid as ACCESS-S using bilinear interpolation as a first step. In a second step, the original ADFD forecast for the AWS location was then re-imposed on the nearest grid cell to the AWS.

For each location and validity period, twelve pieces of daily precipitation forecast information are published to the ADFD: four quantile forecasts[3], seven PoE forecasts and the mean value forecast (see Table 1). The PoE forecasts are published to the nearest 0.01. The PoE forecast for the threshold 0 mm is the probability of precipitation (PoP) forecast. Daily Tmin and Tmax published to the ADFD are mean-value forecasts. No other probabilistic information for daily temperature is published. Validity periods for ADFD daily precipitation, Tmin and Tmax forecasts match those of the observations. The ADFD dataset used in this report for verification had fewer than 1% of forecasts missing across lead days 1 to 7.

4.4 Lead day conventions

Each day, ADFD forecasts are published in the late afternoon and the processing of the ACCESS-S forecast is complete in the early evening. 'Day 1' forecasts for daily precipitation, Tmin and Tmax are the forecasts that are valid for the following local day, and 'Day 2' forecasts are those that are valid for the local day after that. That is, lead day is relative to the time when the forecasts are available for use (the 'availability time').

Table 2 illustrates the relationship between the 'base time', availability time and validity periods for lead day 1 forecasts. For ACCESS-S, the base time is the model initial condition time for the most recent 33-member ensemble used in the time-lagged ensemble. For the ADFD, the base time is the initial condition time for the most recent global NWP models that are typically used as inputs for the ADFD.

4.5 Verification methods and reference forecasts

Predictive accuracy of the various ensemble forecasts is assessed against observations using the continuous ranked probability score (CRPS), which is a commonly-used strictly proper

[3]A quantile forecast at the 0.25 level, or a 0.25-quantile forecast, is the same as a 25th percentile forecast. If the 0.25-quantile forecast is 10 mm, this is equivalent to a forecasting a 25% chance that the precipitation will be 10 mm or less.

Table 2: The relationship between base time, availability time, forecast valid end time and forecast duration (in hours) for lead day 1 forecasts that are valid for the local date of 3rd January 2024. All times are in UTC.

system	base time	availability time	variable	day 1 valid end	duration
ADFD	2024-01-01 12:00	2024-01-02 07:00	precip	2024-01-03 15:00	24
			Tmax	2024-01-03 13:00	18
			Tmin	2024-01-03 01:00	18
ACCESS-S	2024-01-01 00:00	2024-01-02 09:00	precip	2024-01-03 15:00	24
			Tmax	2024-01-04 00:00	24
			Tmin	2024-01-03 00:00	24

scoring rule for measuring the accuracy of predictive distributions of a single real-valued variable (Matheson and Winkler, 1976; Gneiting and Raftery, 2007). The CRPS can be interpreted as a generalization of the absolute error for single-valued forecasts, and the units are the same as the observations. There are several methods of calculating the CRPS for an ensemble. We use the 'fair' method of Ferro et al. (2008), where for a single forecast case we have

$$\mathrm{CRPS}(\mathbf{x}, y) = \frac{1}{n}\sum_{i=1}^{n}|x_i - y| - \frac{1}{2n(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{n}|x_i - x_j|, \tag{4}$$

where $\mathbf{x} = (x_i)_{i=1}^{n}$ is the n-member ensemble, x_i is the value of the ith ensemble member and y is the corresponding observation. The CRPS is negatively orientated. A score of zero indicates a perfect forecast (i.e., every ensemble member forecast the observed value). The first summation in Equation (4) measures the average distance between the ensemble members and the observation, while the second set of summations adjusts the penalty based on ensemble spread.

A single-valued forecast x can be interpreted as a deterministic distribution, in which case $\mathrm{CRPS}(x, y) = |x - y|$. This will be used to compare ensemble prediction with some single-valued forecasts from the ADFD.

Results in this report are presented as skill scores with respect to a reference forecast. The continuous ranked probability skill score (CRPSS) for a forecast system A over a set of forecast cases is calculated by

$$\mathrm{CRPSS} = 1 - \overline{\mathrm{CRPS}}_{\mathrm{A}}/\overline{\mathrm{CRPS}}_{\mathrm{ref}}\,,$$

where $\overline{\mathrm{CRPS}}_{\mathrm{A}}$ is the mean CRPS for System A over the set of forecast cases, and $\overline{\mathrm{CRPS}}_{\mathrm{ref}}$ is the mean CRPS for a reference forecast system over the corresponding set of forecast cases. A CRPSS of 1 indicates every forecast was perfect, while a positive CRPSS indicates that System A had more accurate predictions on average than the reference forecast. A negative CRPSS indicates that the reference forecast was more accurate on average than System A.

Reference forecasts for 24-hour precipitation are constructed from a 38-year (1981 to 2018) gridded analysis of daily rainfall data from the Australian Gridded Climate Dataset

(AGCD) (Jones et al., 2009). For each location, all analysis values within 15 days of that day of year are selected to form an 'ensemble of 24-hour accumulations' of size 1178 (31×38) as the reference forecast. For 7-day precipitation accumulations, the reference forecast is constructed in the same way using 7-day accumulations from AGCD. Reference forecasts for daily Tmin and Tmax are constructed similarly, but use observations at each AWS for the 10-year period 2012 to 2021 rather than analyses from AGCD. Thus the reference ensemble will have a maximum size of 310 observations (31×10), but due to missing observational data may sometimes have less. We required at least 165 observations to form a reference forecast for a given location and day of the year. Note that reference forecasts are constructed from observations or analyses for periods prior to the verification periods.

Tests of statistical significance for the difference in predictive skill between two forecast systems are based on the Deibold–Mariano test for equipredictive performance (Diebold and Mariano, 2002). For a fixed lead day, the difference in CRPS between two forecast systems is calculated for each location and validity period. Differences are then averaged spatially, thus removing spatial dependence, to construct a time series of mean CRPS differences. The Hering and Genton (2011) modification of the Deibold–Mariano test statistic, which accounts for temporal dependence, is then calculated from the time series.

Prior to any evaluation, missing data in one forecast source is removed from other forecast sources, so that skill scores are calculated using a common set of forecast cases. Overall, this resulted in the elimination of about 20% of all possible forecasts cases from the verification periods of this study.

5 Application of Case 1 methods to Bureau precipitation forecasts

5.1 Set up

In this section we illustrate the two seamless transformations ϕ_s and ϕ_w of Section 2.2 using daily precipitation forecasts from the Bureau. The ensemble forecasts come from the 99-member ACCESS-S ensemble while the target predictive distribution is derived from the set of precipitation forecasts published to the ADFD (see Table 1). The methods in this section will be illustrated using lead day 2 forecasts for daily precipitation at Moree Airport for 21 December 2023 (location indicated in Fig. 1, and hereafter referenced as 'Moree forecasts').

In order to generate the seamless transformations ϕ_s and ϕ_w, detailed information about the ADFD precipitation quantile function F^{-1} for each location and validity period is obtained via a reconstruction process, as illustrated in Fig. 6. We begin with the seven PoE forecasts and four quantile forecasts listed in Table 1, which give up to eleven values of F^{-1} (the green circles and blue squares on Fig. 6). For example, the green circle furthest to the right in Fig. 6 corresponds to the ADFD PoE forecast $\mathbb{P}(Y > 50) = 0.08$. In order to gain more information about the tail of the distribution, we extrapolate to obtain the 95th to 99th percentiles (dark orange diamond markers) using the Weibull distribution, as detailed in Appendix C. The remainder of F^{-1} is then reconstructed via linear interpolation and extrapolation. In most forecast cases, the Weibull extrapolation step is not required because $\mathbb{P}(Y > 50) \leq 0.01$.

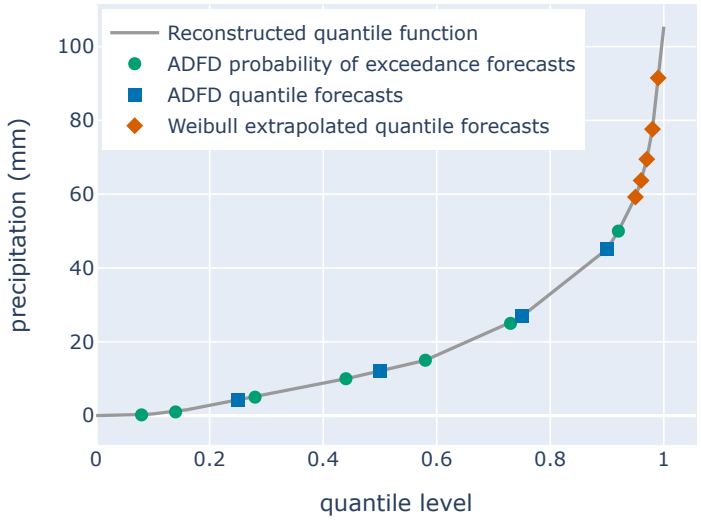

Figure 6: Graph (grey line) of the reconstructed ADFD lead day 2 daily precipitation quantile forecast F^{-1} at Moree Airport valid 21 December 2023. The reconstruction uses ADFD PoE forecasts (green circles), ADFD quantile forecasts (blue squares) and quantiles extrapolated using the Weibull distribution (dark orange diamonds).

The strongly consistent transformation ϕ_s requires a ranking function R. In this case, at each grid cell the ensemble members are ranked first by precipitation value (a higher value corresponds to higher rank). If there are any ties, these are broken first by ranking according to the forecast ACCESS-S Tmax value at the grid cell (a lower value corresponds to higher rank, where the Tmax forecast taken is +9 hours relative to the precipitation forecast) and finally using a random ranking. The random ranking is generated for each validity period and model run, but is otherwise constant across the grid, to avoid noisy spatial output for each seamless ensemble member. Other candidate ranking functions and their impact on spatial coherence and predictive accuracy are discussed in Section 5.3.

The weakly consistent transformation ϕ_w requires an algorithm $\mathcal{F}$ for fitting the distribution of ensemble members $\mathbf{x}$ to a parameterized distribution. We begin with a modified ensemble $\tilde{\mathbf{x}}$ obtained by mapping all ensemble members with a value less than 0.2 mm to the value 0 mm. This step removes many cases of small positive values from the ensemble which are below the resolution of the observations and can have a negative impact on curve fitting. In each forecast case, a 3-parameter gamma distribution is fitted to the set of positive ensemble members from $\tilde{\mathbf{x}}$ using a method described in Appendix A. This is then combined with the probability of dry conditions, as detailed in Appendix A, to obtain the final parametrized fit $\mathcal{F}(\tilde{\mathbf{x}})$. An example of a fit $\mathcal{F}(\tilde{\mathbf{x}})$, which we denote by G, is the green curve of Fig. 7 for the modified ACCESS-S forecast $\tilde{\mathbf{x}}$ at Moree Airport. In

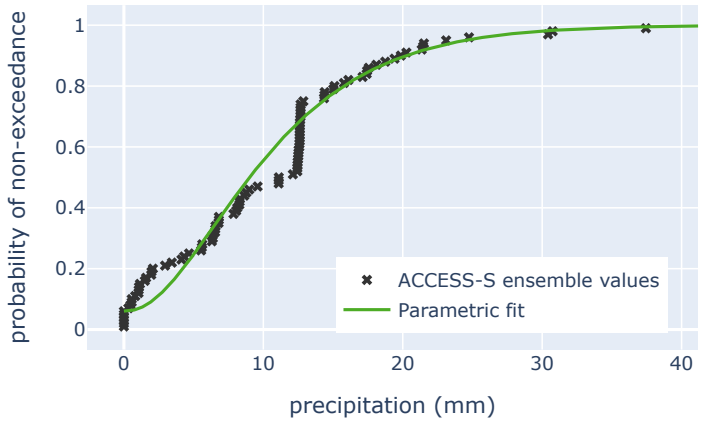

Figure 7: Parametric fit (green curve) using a gamma distribution to values from the 99-member ACCESS-S ensemble (dark crosses) lead day 2 precipitation forecast for Moree Airport valid 21 December 2023.

this case, 6 out of 99 modified ensemble members are 0 mm so $G(0) = 6/99$.

In practice, we make a small modification to the weakly transformed ensemble $\phi_{\mathrm{w}}(\tilde{\mathbf{x}}, F; \mathcal{F})$ so that the ensemble PoP forecast is strongly consistent with the ADFD PoP forecast. Specifically, we first rank members of the original ensemble using R. Then for each forecast case the ADFD PoP is $k/100$ for some integer k. Any transformed ensemble member whose original rank was lower than k is set to 0 mm, and any transformed ensemble member with a value less than 0.2 mm whose original rank was at least k is set to 0.2 mm.

With this choice of R and $\mathcal{F}$, we obtain a strongly and a weakly consistent seamless transformations of the original ensemble. Call these the Seamless-S and Seamless-W forecasts. They are illustrated for the Moree example in Fig. 8.

5.2 Objective evaluation

We now evaluate the Seamless-S and Seamless-W forecasts against four criteria: (1) consistency with published ADFD probabilistic forecasts, (2) predictive accuracy of day 1 to 7 forecasts, (3) predictive accuracy of the week 1 (168-hour) accumulation forecasts and (4) computation time. Evaluation is performed using almost 12 months of data at 522 locations across Australia using the precipitation datasets described in Section 4.

The level of consistency of each of the ensemble forecasts (ACCESS-S, Seamless-S and Seamless-W) with the ADFD was measured using the mean absolute difference (MAD) of all forecast cases in the verification period. Fig. 9 shows the MAD between the ADFD 0.75-quantile forecast and the corresponding 0.75-quantile forecast for ACCESS-S and the two seamless ensembles. A lower MAD indicates closer agreement between ADFD and the ensemble. Unsurprisingly, Seamless-S had the lowest MAD due to strong consistency (as

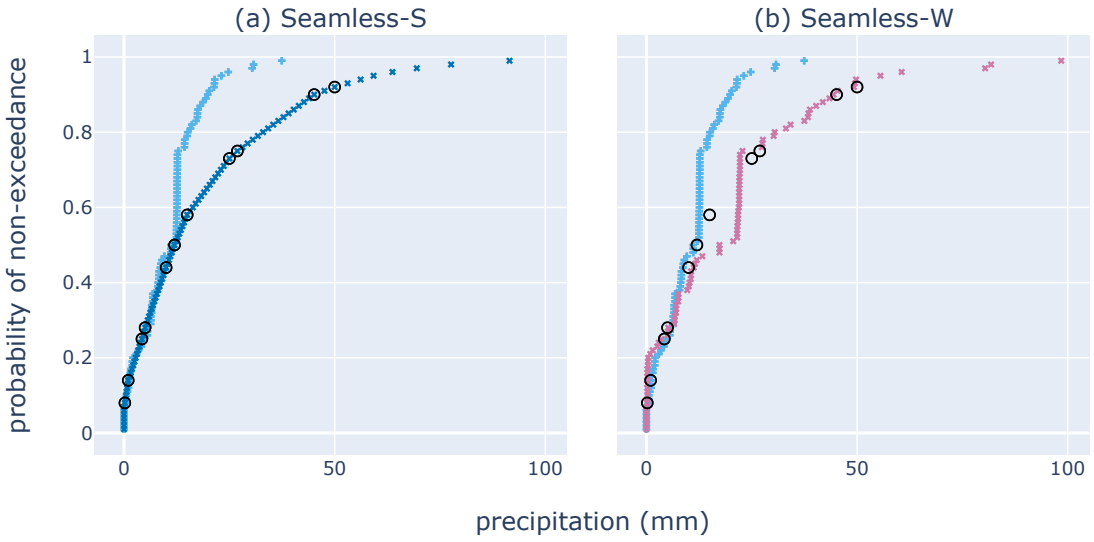

Figure 8: Lead day 2 forecasts for Moree Airport valid for 21 December 2023 from (a) Seamless-S (dark blue × markers) and (b) Seamless-W (pink × markers), compared with the ACCESS-S ensemble forecast (light blue + markers). The black open circles are the published values from the ADFD forecast.

exemplified by the alignment of dark blue crosses with black circles in Fig. 8), but both seamless forecasts had substantially higher agreement with the ADFD than ACCESS-S. The results were similar for other quantile and PoE forecasts (not shown). Full consistency between Seamless-S and ADFD is not possible because of rounding and occasional internal inconsistencies in the ADFD forecasts.

The predictive accuracy of the day 1 to 7 ensemble forecasts when compared to observations, as measured by CRPSS, is shown in Fig. 10a. Both seamless ensemble forecasts, and by implication the ADFD forecasts on which they are based, are substantially more skillful than ACCESS-S for 24-hour accumulation over all lead days. Seamless-S is also slightly more accurate than Seamless-W, which is statistically significant at the 5% level for all lead days.

By summing daily precipitation for each ensemble member across the 7 days, week 1 (168-hour) forecast accumulations are obtained. The only ADFD week 1 accumulation that can be readily derived from ADFD daily forecasts is the week 1 mean value, which is the sum of the daily mean forecasts.[4] In this evaluation, the ADFD week 1 mean value forecast, being a single value, will be interpreted as a deterministic distribution. Fig. 10b shows the CRPSS for week 1 precipitation for the ADFD mean and each of the ensembles. Each of the

[4]It is not possible to reconstruct ADFD week 1 quantiles from daily ADFD quantiles or PoE forecasts, though it is possible to infer some bounds. For example, the 0.9-quantile for week 1 is at least as large the maximum of the daily 0.9-quantiles.

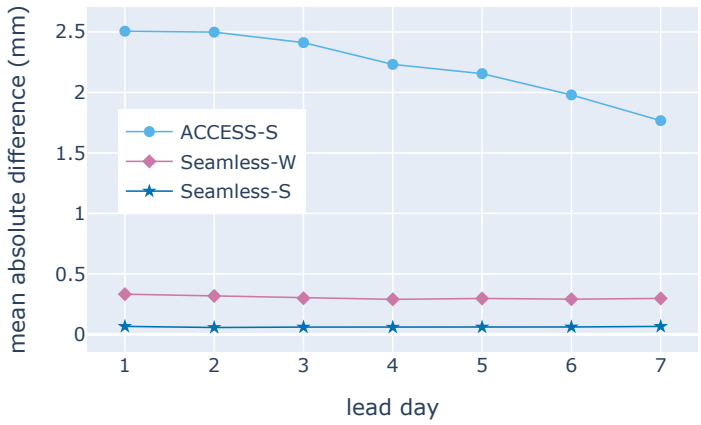

Figure 9: Mean absolute difference (MAD) by lead day between ADFD 0.75-quantile forecasts for daily precipitation and corresponding forecasts from three different ensembles. A lower MAD indicates closer agreement between ADFD and the ensemble.

seamless week 1 forecasts is substantially more skillful than ACCESS-S, while the single-valued ADFD forecast is slightly less accurate than the climatological reference forecast distribution. Interestingly, while Seamless-S was slightly more accurate than Seamless-W for daily forecasts, this advantage has almost been erased for the week 1 forecasts, and the difference in predictive accuracy between the two seamless week 1 forecasts is not statistically significant at the 5% level.

The predictive skill of the ensemble mean can be measured *consistently* by mean squared error (Gneiting, 2011) and reported by the root mean squared error (RMSE). The RMSEs of week 1 Seamless-S, Seamless-W and ADFD mean value forecasts were very close at 21.93 mm, 21.88 mm and 21.96 mm respectively, and the null hypotheses of equipredictive performance are accepted the 5% significance level using Diebold–Mariano test statistics. Note that ADFD mean value forecasts are not used to construct the seamless forecasts. The RMSE for the ACCESS-S ensemble mean was 28.97 mm.

Seamless-S is substantially faster to compute than Seamless-W. There are two main reasons. First, the Seamless-S algorithm does not require curve fitting whereas the Seamless-W algorithm does. The second and more significant reason is that the Seamless-S algorithm computes values of the ADFD quantile function F^{-1} for a fixed set of quantile levels based on ensemble size (in our case, levels $0.01, 0.02, \ldots, 0.99$), whereas the Seamless-W algorithm calculates values of F^{-1} at many different quantile levels. This is computationally intensive over larger gridded data sets since F^{-1} is piecewise linear with join points varying across the domain. Were F^{-1} one of many standard parametric quantile functions this would not be a problem. Nonetheless, employing the weakly consistent seamless algorithm in Bureau operations would only increase processing time by several

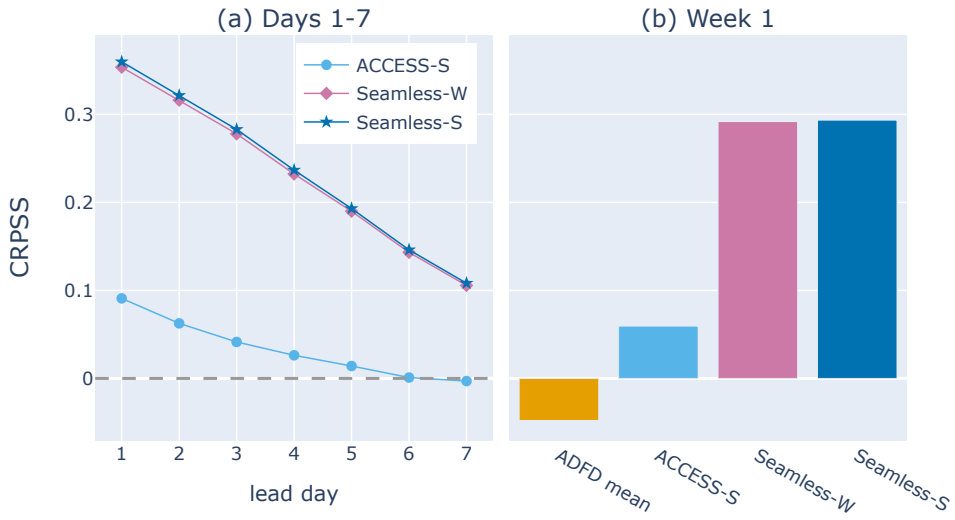

Figure 10: Continuous ranked probability skill scores for (a) day 1 to 7 and (b) week 1 precipitation forecasts. A higher skill score is better.

minutes compared to using the strongly consistent seamless algorithm, and does not pose a major barrier for adoption.

Overall, Seamless-S ranks as the best method on computation efficiency, consistency with ADFD and predictive accuracy. This is the method that the authors have implemented as a post-processing step to ACCESS-S in the Bureau's research production pipeline.

5.3 Subjective evaluation of spatial outputs and choice of ranking function

The spatial coherence of Seamless-S forecasts is important for two reasons. First, many of the Bureau's multi-day statistical forecast products are presented spatially (e.g. Fig. 1a) and any obvious spatial incoherence tends to erode user trust in those products. Second, some users view spatial outputs from individual ensemble members, each of which is interpreted as one possible plausible scenario. Fig. 11 gives examples of spatial outputs of 24-hour and 7-day accumulations from ACCESS-S, Seamless-S and ADFD. The Seamless-S and ADFD 7-day mean accumulations (row 1) match each other exceptionally well, noting that the ADFD mean is never used to create the seamless forecast. In contrast, not only is the ACCESS-S ensemble mean (Fig. 11a) noticeably lower over most of the domain but it also contains spatial artifacts, namely the two regions of no rainfall over the western interior, which is due to the calibration scheme over regions with sparse observations.

The 24-hour forecasts (rows 3 and 4 of Fig. 11) deserve special discussion because they give an example of a common problem encountered in recalibration approaches: how

Table 3: CRPSS of week 1 precipitation forecasts. Each of the Seamless-S forecasts is strongly consistent with the ADFD, but generated from different ranking functions R, R_1 and R_2 as indicated in brackets. A higher CRPSS is better.

Forecast	CRPSS
ACCESS-S	0.0598
Seamless-S[R]	0.2936
Seamless-S[R_1]	0.2968
Seamless-S[R_2]	0.2103

to assign rainfall to some ensemble members where there is no rainfall in the original ensemble. The synoptic situation is a frontal rain-bearing cloud band extending from the northwest to the southeast. The ensemble means (third row) show that on average the front lies further to the west with ACCESS-S than with ADFD and Seamless-S, so that ADFD has rainfall over parts of the northern interior where ACCESS-S has none. To be consistent with ADFD rainfall, some ensemble members in Seamless-S must be assigned rainfall in these parts where ACCESS-S is completely dry. In this situation, ensemble members from ACCESS-S are ranked by Tmax at each grid cell, since first ranking by precipitation where all ACCESS-S members forecast 0 mm results in ties. The 6th ensemble member (bottom row) of Seamless-S is one such ensemble member that is assigned rainfall, with falls exceeding 20mm over this part of the northern interior (Fig. 11k). Even though such rainfall is the creation of the seamless algorithm, it looks no less spatially 'realistic' than ensemble member 6 from ACCESS-S (Fig. 11j).

Recall that the only information used from ACCESS-S to create Seamless-S is the rank of ensemble members via the ranking function R (ranking first by ACCESS-S precipitation then by ACCESS-S Tmax then randomly). To assess the value of this information, consider Seamless-S[R_1] and Seamless-S[R_2] forecasts which are generated using alternative ranking functions R_1 and R_2 that use no information from ACCESS-S. The function R_1 assigns a random rank to each ensemble member for each validity period and model run; the rank does not vary spatially. The function R_2 assigns a rank equal to the ensemble member number, with the result that ensemble member k from Seamless-S[R_2] is the kth percentile forecast from the original ensemble. For clarity, denote the original Seamless-S forecast by Seamless-S[R] for the remainder of this section. The accuracy of week 1 precipitation forecasts for the precipitation verification period across Australia, as measured by the CRPSS, is presented in Table 3. Each of the seamless forecasts generated by ϕ_s (with different ranking functions) performed better than ACCESS-S, but somewhat intriguing is that seamless forecasts generated from the random ranking function R_1 performed best over this dataset, outperforming the seamless forecast with ranking function R at the 10% statistical significance level, but not at the 5% level. The experiment was repeated several times with similar results. Seamless-S[R_1] is also computationally cheap because it does not require a dynamical model ensemble. See Section 7 for further discussion.

However, if users would like output from individual ensemble members to have the spatial appearance of a possible plausible scenario, Seamless-S[R] outputs are more accept-

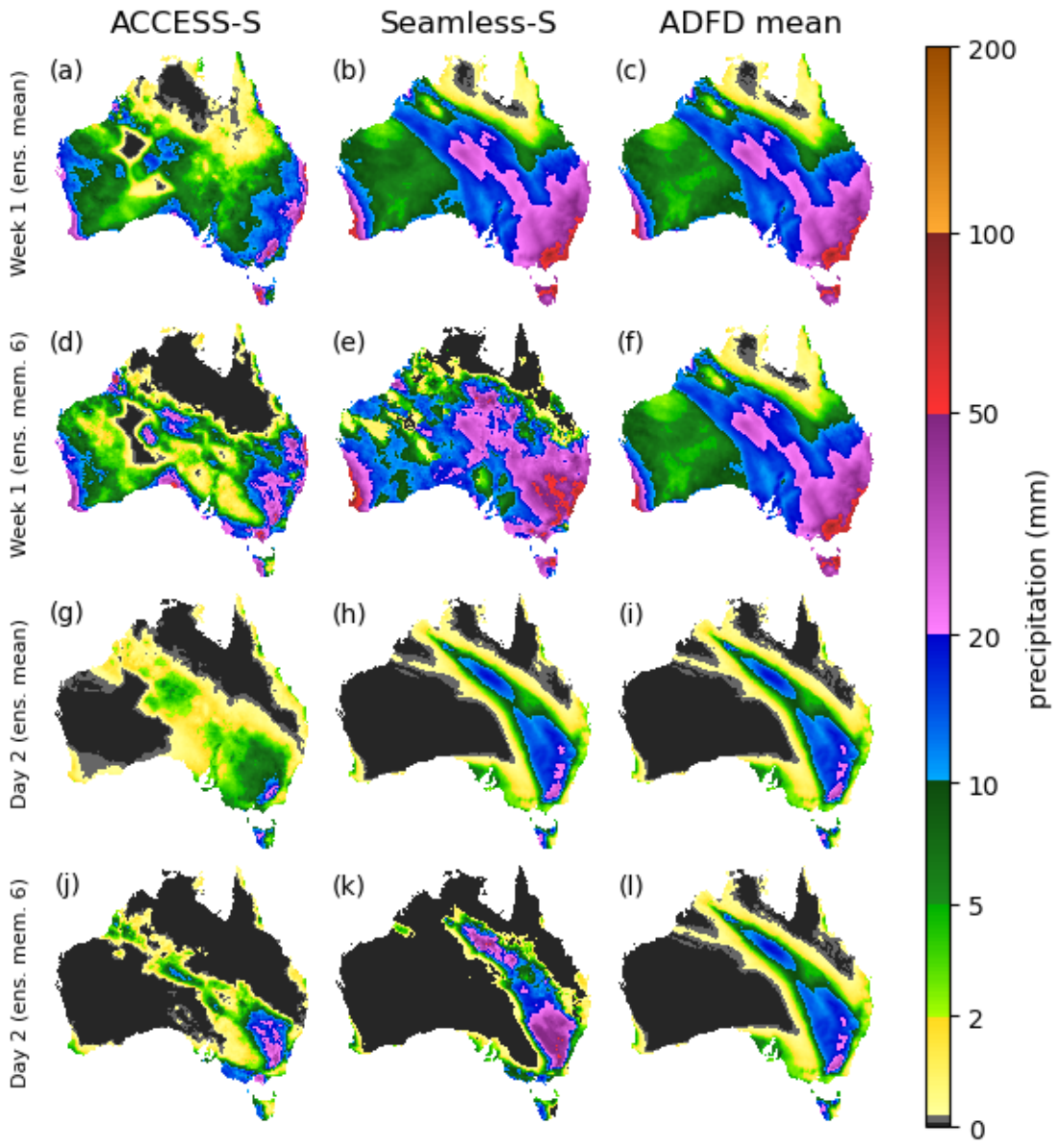

Figure 11: Precipitation forecasts from ACCESS-S (column 1), Seamless-S (column 2) and the ADFD mean value (column 3), for the 7-day accumulation week 1 forecast for the period 30 May to 5 June 2024 (rows 1 and 2) and the 24-hour Day 2 forecast for 31 May 2024 (rows 3 and 4). For the ensemble forecasts, rows 1 and 3 are the ensemble mean while rows 2 and 4 are for ensemble member 6.

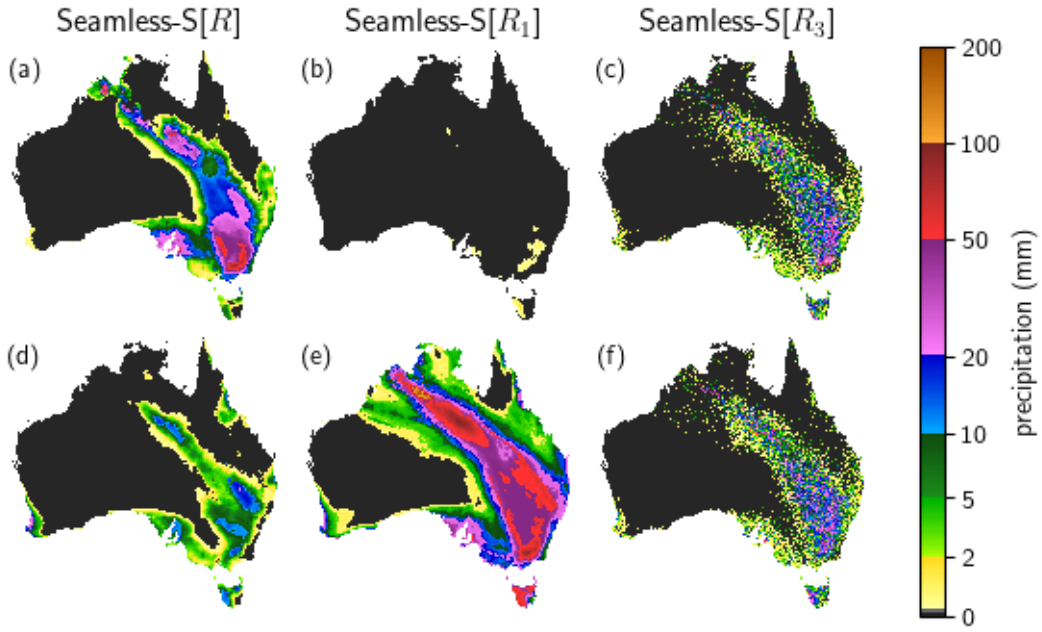

Figure 12: 24-hour accumulation forecasts of precipitation for 31 May 2024 for selected ensemble members, using the strongly consistent seamless algorithm with three different ranking functions R, R_1 and R_3. The selected Seamless-S[R] ensemble members were those that produced highest accumulation over (a) the southeastern ranges and (d) the far southwest. The selected Seamless-S[R_1] ensemble members were those that ranked (b) lowest and (e) highest. The two Seamless-S[R_3] ensemble members (c and f) were randomly chosen.

able than outputs from Seamless-S[R_1]. Fig. 12 shows two ensemble members from each of Seamless-S[R], Seamless-S[R_1] and Seamless-S[R_3], where R_3 ranks ensemble members randomly for each grid cell rather than uniformly across the domain. The two ensemble members from Seamless-S[R] were chosen as they gave highest totals over the southeastern ranges (Fig. 12a) and far southwest (Fig. 12d). Neither gave highest totals at each grid cell in the domain. The two ensemble members from Seamless-S[R_1] were from the ensemble members that were ranked lowest (Fig. 12b) and highest (Fig. 12e). Neither of these Seamless-S[R_1] ensemble members represents a physically plausible possible scenario given the particular synoptic situation, with the first producing a dry scenario over most of the continent in the presence of a frontal rain band, and the second which would lead to widespread flooding over much of the interior. Rather, they should be interpreted statistically as the 1st and 99th percentile forecasts for each grid cell. Seamless-S[R_3] (Figs. 12c and f) has no spatial coherence at small scales where rain is possible.

6 Application of Case 2 methods to Bureau temperature and precipitation forecasts

6.1 Application to Bureau temperature forecasts

We apply the transformation ϕ_{b} of Section 3 to generate Seamless-B Tmin and Tmax forecasts from the 99-member ACCESS-S ensemble at lead days 1 to 7 so that they are consistent with ADFD mean value forecasts. The accuracy of ACCESS-S, Seamless-B and ADFD forecasts are assessed using the 12-month Australia-wide temperature datasets detailed in Section 4. Recall that the ADFD grid cell elevation matches the AWS elevation. On the other hand, the AGCD gridded reanalysis, on which ACCESS-S is trained, does not. To obtain a fair comparison, we also create a new ensemble called 'ACCESS-S debiased' by calculating the bias between AWS observations and the AGCD reanalysis for the 3-year period 2016-2018 and subtracting that bias from ACCESS-S. For most sites the bias was not substantial, but for some AWSs in mountainous terrain bias corrections in the order of 5°C were made.

The bound parameters ℓ and u for each AWS location are based on diabiased AGCD climatology for the period 1981-2018 with additional universal bounds. Climate extremes at each location are calculated using extreme value analysis to obtain thresholds ℓ_1 and u_1 which are 1-in-100 year events.[5] Then we set the lower threshold $\ell = \max(\ell_1 - 4, -25)$ and the upper threshold $u = \min(u_1 + 4, 54)$. For the lower bound, subtracting 4°C from ℓ_1 allows for cold events at the location which are more extreme than the 1-in-100 year threshold, while applying a universal cap of −25°C gives a global lower bound for temperature in Australia. The global upper bound selected is 54°C, though this could be revised upward in the future based on warming climate trends. The buffer parameters selected were $b_1 = b_2 = 1$.

The forecast accuracy of ADFD, ACCESS-S, debiased ACCESS-S and Seamless-B for lead days 1 to 7 is shown in Figs. 13a and 13c. Here, the single-valued ADFD forecast is interpreted as a deterministic distribution. Debiasing ACCESS-S leads to some skill gain. Despite the ADFD being a single-valued forecast, it outperformed debiased ACCESS-S Tmax forecasts at days 1 to 5. Nonetheless, Seamless-B performed best by combining the skill of ADFD with uncertainty information from ACCESS-S. All these forecasts had positive skill relative to the climatological reference forecast, apart from ADFD at days 6 and 7 for Tmin whose single-valued forecast is competing with a reference distribution.

The mean minimum (maximum) temperature forecast for week 1 is the average of the daily Tmin (Tmax) forecasts across the 7 days. The relative accuracy of week 1 forecasts for each ensemble and the ADFD is then assessed using CRPSS, with results shown in Figs. 13b and 13d. For both Tmin and Tmax, ADFD and the ensembles perform better than the reference forecast, while Seamless-B performs best overall. This shows the value of combining uncertainty information from an ensemble (ACCESS-S) along with a well-calibrated mean-value forecast (ADFD) to create a superior seamless product.

[5] Time series of block maxima or minima were formed for each station using debiased AGCD climatology (1981-2018) for each station. Each block was 12 months long, starting in July for maxima and January for minima, yielding time series of length 37 for maxima and of length 38 for minima. Each time series was fit to a generalized extreme value distribution using the method of probability weighted moments (Hosking et al., 1985). Thresholds ℓ_1 and u_1 were obtained from probabilities of the fitted distributions, i.e., $\mathbb{P}(T < \ell_1) = 0.01$ and $\mathbb{P}(T > u_1) = 0.01$, where T is either Tmax or Tmin as appropriate.

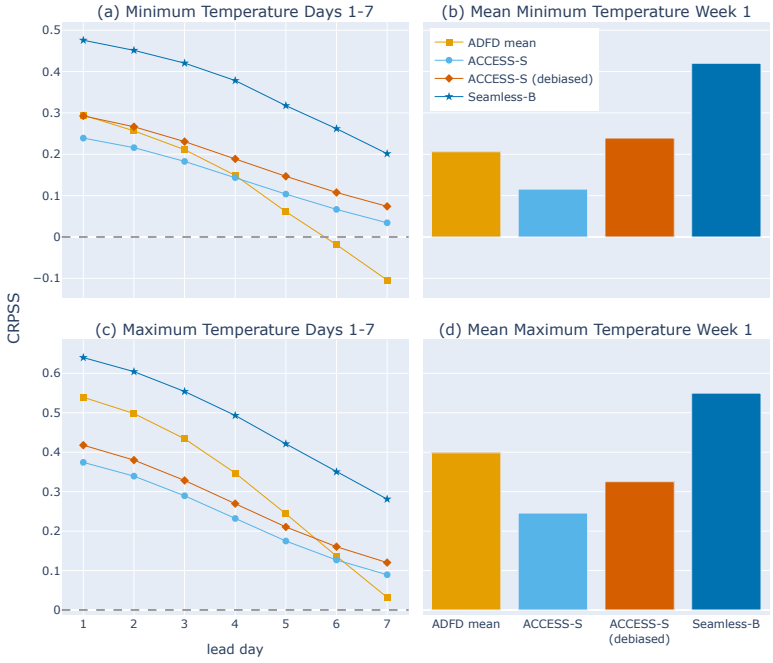

Figure 13: Continuous ranked probability skill scores (CRPSS) for daily and week 1 mean Tmin and Tmax forecasts, Australia wide, for the 12-month period starting 11 November 2022. A higher score is better.

6.2 Application to precipitation forecasts

The transformation ϕ_{b} could also be used to transform an ensemble of precipitation or wind speed forecasts to match the mean value of a target distribution. In this case, a sensible choice for the lower bound ℓ is 0 (mm or ms^{-1}) and for the lower buffer b_1 is 0. We apply this method to daily precipitation forecasts $\mathbf{x}$ from the ACCESS-S ensemble, using the ADFD daily mean precipitation forecast as the target z, and 2000 mm as the upper bound u and setting $b_2 = 0$. With this choice of bound parameters, ϕ_{b} generates Seamless-B precipitation forecasts. Fig. 14 shows the accuracy of predictive distributions of daily precipitation, as measured by CRPSS, for Seamless-B, ACCESS-S, Seamless-S, ADFD (mean value only, treated as a deterministic distribution) and the climatological reference forecast using the precipitation dataset of Section 4. The Seamless-B ensemble outperforms the ACCESS-S ensemble for lead days 1 to 6. Seamless-S predictive distributions were most accurate on average, noting that Seamless-S uses much more information from the ADFD distribution than is contained in the ADFD mean.

We do not claim that ϕ_{b} is a good transformation for precipitation and wind speed forecasts, but it could be used a benchmark to compare for other transformations that aim to output an ensemble that is consistent with the mean value from some target forecast.

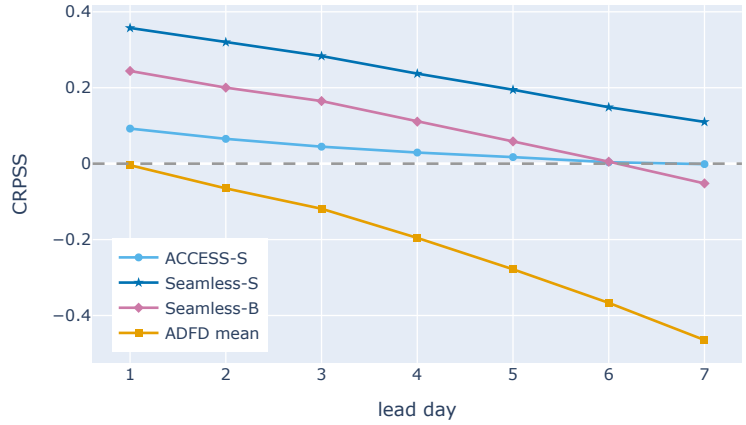

Figure 14: CRPSS for predictive distributions of daily precipitation. Higher values are better.

7 Discussion and conclusions

We have described several methods for adjusting forecasts from an ensemble so that they are statistically consistent with either predictive distributions or mean value forecasts from a second (target) forecast system. One demonstrated use case is modifying an ensemble that is designed for subseasonal to seasonal forecasting so that it is consistent, in the first 7 days, with weather forecasts from another system. The original motivation for this work was to provide users an offering of more consistent forecasts across different forecast horizons, which increases user confidence in, and their ability to correctly interpret, the products. When implemented, we also found that these methods improved predictive skill of the ensemble from days 1 to 7. A strength of ensemble forecasts is that aggregating the forecast over multiple time steps is easy. Consequently, improved day 1 to 7 predictions lead to improved multi-day predictions that intersect with the day 1 to 7 period. For example, we found that week 1 Seamless-S precipitation and week 1 Seamless-B temperature forecasts were substantially more accurate than week 1 forecasts from the original ensemble (ACCESS-S) and from the weather forecasts (ADFD), from which these seamless forecasts were constructed. These methods are computationally cheap, relative to improving ensemble forecasts through increased model resolution, improved model physics and numerics, or implementing different calibration schemes for different forecast horizons.

For adjusting an ensemble to be consistent with a target predictive distribution, the strongly consistent transformation ϕ_s (Equation (1)) and the weakly consistent transformation ϕ_w (Equation (2)) are two viable options. If close agreement between the transformed ensemble and the target distribution is a premium, then ϕ_s is recommended. On the other hand, if preserving the character of the original ensemble members (e.g. that the 5th ensemble member is an outlier scenario) is important, then ϕ_w is recommended. For this reason, we also hypothesize that ϕ_w may be more appropriate if the character of

the combined narrative across multiple forecast fields (e.g., temperature, wind and rain) for ensemble members is to be preserved to some extent. Following initial investigation reported in Section 5.2, further work and data is required to determine whether there are predictive advantages of using one of the transformations over the other when it comes to aggregating forecasts over multiple time steps.

An important feature of the strongly consistent transformation ϕ_s is that it could be used to produce an ensemble of forecasts that is consistent with target predictive distributions, without input from some pre-existing ensemble. The target distribution and size of the ensemble alone determine the set of values within the ensemble. The critical decision (determined by the choice of ranking function R) is how those values are mapped to each of the ensemble members. In Section 5.3 we found that assigning the values to ensemble members at random produced marginally more accurate aggregated precipitation forecasts than assigning the values to ensemble members based on their ranking within the original ACCESS-S ensemble. Possible explanations as to why using information from ACCESS-S produces marginally less accurate multi-day predictions is the subject of further exploration. For example, repeating the study on the 33-member ACCESS-S ensemble, rather than the 99-member time-lagged ensemble, may yield different results, particularly if time-lagging degrades the forecast at short lead times. Exploring accuracy of different methods for different rainfall climates or regimes may also provide insight. On the other hand, it may be that ϕ_s, with random ranking, is actually difficult to beat for predictions of week 1 precipitation over Australia as a whole. If so, this report provides a computationally cheap method for generating multi-day predictive distributions of precipitation from day 1 to 7 weather forecasts produced by the Bureau.

For adjusting an ensemble to be consistent with a target mean-value forecast, the transformation ϕ_b (Equation (3)) was found to be a suitable option for temperature predictions, as it generated both consistent forecasts and substantially improved predictive performance. For forecast variables with a hard attainable lower bound (such as 0 mm for rainfall or 0 ms^{-1} for wind speed) ϕ_b achieves consistency but it is likely that other consistent transformations will result in substantially more accurate prediction. The transformation ϕ_b was built around shifting forecasts by a suitable constant; perhaps a better transformation for wind speed would be built around adjusting forecasts by a suitable scale factor, or using some other transformation that works well with hard lower bounds.

In summary, this report has presented an exploration of the statistical methods that may be used to generate a single set of ensemble forecasts that combine the best information from separate systems across the timescales of the daily weather forecast and into the subseasonal range. One of these methods (Seamless-S) has already been implemented in the Bureau's research production pipeline for precipitation, whilst another (Seamless-B) is currently in preparation for the research production pipeline.

Acknowledgements

Initial aspects of this work were funded by the Agri-Climate Outlooks (ACO) project of Agriculture Innovation Australia. Later aspects of this work were partly funded by Meat & Livestock Australia, the Queensland Government through the Drought and Climate Adaptation Program, and the University of Southern Queensland through the Northern Australia Climate Program (NACP). The authors are grateful to Robin Wedd who as-

sisted making 15 UTC-aligned daily precipitation forecasts from ACCESS-S available, and to Deryn Griffiths and Beth Ebert for providing feedback on an earlier version of this manuscript.

Appendices

We establish some notation that will be used in the appendices. Recall that $\mathbf{x}$ denotes a tuple $(x_i)_{i=1}^n$ of ensemble forecasts, and that $x_{(j)}$ denotes the jth smallest ensemble member. Define its mean $\bar{\mathbf{x}}$, variance $\mathrm{var}(\mathbf{x})$ and standard deviation $\mathrm{sd}(\mathbf{x})$ by

$$\begin{aligned}\bar{\mathbf{x}} &= \frac{1}{n}\sum_{i=1}^{n} x_i,\\ \mathrm{var}(\mathbf{x}) &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{\mathbf{x}})^2 \quad \text{and}\\ \mathrm{sd}(\mathbf{x}) &= \sqrt{\mathrm{var}(\mathbf{x})}.\end{aligned}$$

A Parametrized fits to an ensemble of precipitation forecasts

The appendix describes a computationally efficient method to fit parametric distributions to gridded ensemble forecasts of precipitation. In particular, the Seamless-W precipitation forecasts of Section 5 were generated from ACCESS-S ensemble forecasts of precipitation using the curve fitting algorithm $\mathcal{F}$ and the weakly consistent transformation ϕ_{w}. Details of, and background to, $\mathcal{F}$ are given below.

When fitting an ensemble of precipitation forecasts to a parametric distribution, two aspects of the distribution should be accounted for: the probability of precipitation (PoP) and the conditional distribution of precipitation given that precipitation occurs. The conditional distribution has been variously modeled by the 3-parameter gamma distribution (Sloughter et al., 2007; de Burgh-Day and Dillon, 2021), 2-parameter Weibull distribution (some Bureau operational forecasts), the best fitting of gamma or Weibull distributions (de Burgh-Day and Taggart, 2023), or the 4-parameter generalized gamma distribution (Taggart et al., 2022) which includes gamma and Weibull distributions as subfamilies. The algorithm $\mathcal{F}$ of the current report uses computationally efficient methods for estimating gamma distribution parameters, including new methods based on the work of Balakrishnan and Wang (2000) and edge case coverage not implemented by de Burgh-Day and Taggart (2023).

An outline of the curve fitting algorithm $\mathcal{F}$ is as follows. The PoP component is simply the proportion of ensemble members that are wet (i.e., have values greater than 0 mm). Several gamma fits to the nonzero ensemble data are generated using methods (E), (M), (B1) and (B2) of Appendix A.1, to obtain several candidate conditional distributions. Each candidate conditional distribution is then combined with PoP to give a full parametrized distribution. Of the candidate full parametrized distributions, the one that has the lowest 'squared error in probability space' (SEPS) is selected as the final parametric fit. Precise details of the algorithm $\mathcal{F}$ are given in Appendix A.2.

The metric SEPS is equivalent to the mean squared quantile level anomaly (QLA). Given an ensemble forecast $\mathbf{x}$ and a continuous fitted CDF G, $\mathrm{SEPS}(G, \mathbf{x})$ is defined by

$$\mathrm{SEPS}(G, \mathbf{x}) = \frac{1}{n} \sum_{j=1}^{n} \left(G(x_{(j)}) - \frac{j}{n+1} \right)^2 \tag{5}$$

where $x_{(j)}$ is the jth smallest ensemble member. This notion is extended to a discontinuous CDF G by defining the distance d_p in probability space between G and $x_{(j)}$ by

$$d_p(G, x_{(j)}) = \min\{|y - j/(n+1)| : \lim_{z \uparrow x_{(j)}} G(z) \leq y \leq G(x_{(j)})\},$$

whence

$$\mathrm{SEPS}(G, \mathbf{x}) = \frac{1}{n} \sum_{j=1}^{n} d_p(G, x_{(j)})^2. \tag{6}$$

Examples of the application of SEPS are given in Appendices A.2 and B.5.

A.1 Fitting positive ensemble values to a gamma distribution

The CDF F_{g} of the 3-parameter gamma distribution (with real-valued location μ, positive scale σ and positive shape ξ parameters), is given by

$$F_{\mathrm{g}}(x; \mu, \sigma, \xi) = \begin{cases} \dfrac{1}{\Gamma(\xi)} \gamma\left(\xi, \dfrac{x-\mu}{\sigma}\right) & \text{if } x > \mu, \\ 0 & \text{if } x \leq \mu, \end{cases} \tag{7}$$

where γ is the lower incomplete gamma function and Γ is the gamma function. When $\mu = 0$ and $\xi = 1$, F_{g} is the CDF of the standard exponential distribution with rate parameter $1/\sigma$.

Let $\mathbf{w}$ denote the tuple $(w_i)_{i=1}^n$ of non-zero (i.e. wet) ensemble member values for a particular forecast case, ordered such that $0 < w_1 \leq w_2 \leq \ldots \leq w_n$, and assume that this tuple is non-empty. We assume that $\mathbf{w}$ is sampled from a population distributed by a gamma distribution with population parameters (μ, σ, ξ) and the aim is find a tuple $(\hat{\mu}, \hat{\sigma}, \hat{\xi})$ of estimates for these parameters.

Various estimation methods exist but an essential constraint for this particular application is that $0 \leq \hat{\mu} < w_1$, otherwise the transformation ϕ_{w} for producing weakly consistent forecasts is not suitable. As mentioned above, the method should also be vectorizable, which essentially rules out maximum likelihood estimation. We found four methods which, in concert, produced good results whilst meeting these constraints. They are based on a simple fit to the exponential distribution (Method (E)), the method of moments (Method (M)) and two adaptions of a method by Balakrishnan and Wang (2000) (Methods (B1) and (B2)). The latter two methods compute quantities defined by the following equations:

$$m_q(\mathbf{w}, z) = \frac{1}{n} \sum_{i=1}^{n} (w_i - z)^q \tag{8}$$

$$\hat{\xi} = \left| \frac{-0.4 m_1(\mathbf{w}, \hat{\mu}) m_{0.4}(\mathbf{w}, \hat{\mu})}{m_1(\mathbf{w}, \hat{\mu}) m_{0.4}(\mathbf{w}, \hat{\mu}) - m_{1.4}(\mathbf{w}, \hat{\mu})} \right|, \tag{9}$$

$$\hat{\sigma} = (\bar{\mathbf{w}} - \hat{\mu}) / \hat{\xi} \tag{10}$$

Note that when $n = 1$, the denominator of Equation (9) is 0 and $\hat{\xi}$ is undefined. The four methods are described as follows.

(E) Estimates are $\hat{\mu} = 0$, $\hat{\xi} = 1$, $\hat{\sigma} = \bar{\mathbf{w}}$. Uses an unbiased estimator for scale parameter of the exponential distribution. Included as a method because it can be calculated whenever $n \geq 1$.

(M) Estimates are $\hat{\mu} = 0$, $\hat{\sigma} = \text{var}(\mathbf{w})/\bar{\mathbf{w}}$, $\hat{\xi} = \bar{\mathbf{w}}^2/\text{var}(\mathbf{w})$. These estimators are derived beginning with the formulae for the mean and variance of the gamma distribution when $\mu = 0$, then rearranging to make the parameters the subject. Estimates can be computed whenever $n \geq 2$ provided that $\text{var}(\mathbf{w}) > 0$.

(B1) Use the estimate $\hat{\mu} = 0$ and then Equations (8) and (10) to obtain $\hat{\xi}$ and $\hat{\sigma}$. Can be computed whenever $n \geq 2$ and division by 0 is avoided.

(B2) Define δ to be the smallest positive value of the set $\{w_1, w_2 - w_1, w_3 - w_2, w_4 - w_3, w_5 - w_4, w_6 - w_5\}$, excluding any members from that set that don't exist. Then set $\hat{\mu} = w_1 - \delta/2$, so that $0 < \hat{\mu} < w_1$. Thus $\hat{\mu}$ can be calculated whenever $n \geq 1$. Estimate $\hat{\xi}$ and $\hat{\sigma}$ using Equations (9) and (10). Can be computed whenever $n \geq 2$ provided division by 0 is avoided.

Methods (B1) and (B2) start with our own estimate $\hat{\mu}$ and then use method NEW-1 from Balakrishnan and Wang (2000) to calculate $\hat{\xi}$ and $\hat{\sigma}$. This adaption is made because the estimate for μ used by Balakrishnan and Wang (2000), whilst designed to minimize bias, sometimes results in the constraint $0 \leq \hat{\mu} < w_1$ being violated.

A.2 Combining gamma fit with probability of dry conditions

Given an ensemble $(x_i)_{i=1}^n$ of precipitation forecasts for one particular forecast case, let p_{dry} denote the proportion of ensemble members that are dry (i.e. $x_i = 0$). We consider three cases. First, if $p_{\text{dry}} = 1$, so that all ensemble members are dry, then the fitted CDF F is given by

$$F(x) = \begin{cases} 0, & x < 0, \\ 1, & x \geq 0. \end{cases}$$

Second, if $0 < p_{\text{dry}} < 1$ then at one least ensemble member is wet and at least one ensemble member is dry. In this case, the distribution of precipitation conditioned on it occurring naturally satisfies $\hat{\mu} = 0$. We therefore we select the parameter estimates $(0, \hat{\sigma}, \hat{\xi})$ of the best fit, as measured by SEPS, to the wet subset $\mathbf{w}$ of the ensemble using Methods (E), (M) and (B1). Note that at least Method (E) will produce a fit. The final fitted CDF F is then given by

$$F(x) = \begin{cases} 0, & x < 0, \\ p_{\text{dry}} + (1 - p_{\text{dry}}) F_{\text{g}}(x; 0, \hat{\sigma}, \hat{\xi}), & x \geq 0. \end{cases}$$

Third, if $p_{\text{dry}} = 0$, so that all ensemble members are wet, then the final fitted CDF F is given by

$$F(x) = F_{\text{g}}(x; \hat{\mu}, \hat{\sigma}, \hat{\xi}),$$

where $(\hat{\mu}, \hat{\sigma}, \hat{\xi})$ is the best fitting parameter tuple, as measured by SEPS, from among Methods (E), (M), (B1) and (B2).

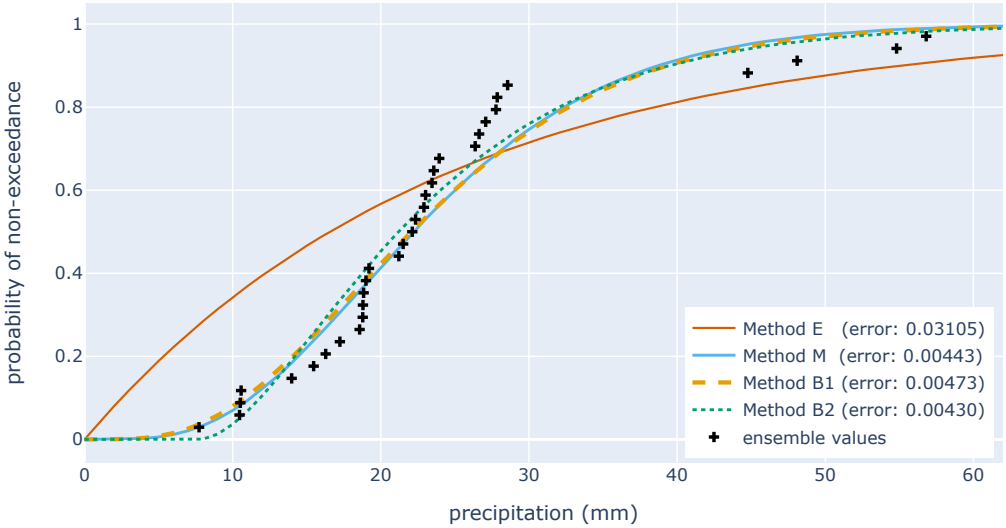

Figure 15: Fitted gamma distributions (curves) to precipitation forecasts from a 33-member ensemble (black crosses), using methods (E), (M), (B1) and (B2). Errors to the fits, as measured by SEPS, are given in parentheses.

These three cases can be combined into one formula using the '3-parameter hybrid gamma distribution' F_{hg} (c.f. de Burgh-Day and Taggart (2023)). Specifically, we introduce a new parameter

$$\hat{\nu} = \begin{cases} -p_{\text{dry}}, & p_{\text{dry}} > 0, \\ \hat{\mu}, & p_{\text{dry}} = 0. \end{cases}$$

That is, if $\hat{\nu}$ is positive it represents (roughly) the smallest precipitation amount that is possible, while if $\hat{\nu}$ is negative its modulus represents the probability of dry conditions. Then the final fitted distribution F is given by

$$F(x) = F_{\text{hg}}(x; \hat{\nu}, \hat{\sigma}, \hat{\xi}) := \begin{cases} F_{\text{g}}(x; \hat{\nu}, \hat{\sigma}, \hat{\xi}) & \text{if } \hat{\nu} \geq 0, \\ |\hat{\nu}| + (1 - |\hat{\nu}|) F_{\text{g}}(x; 0, \hat{\sigma}, \hat{\xi}) & \text{if } -1 \leq \hat{\nu} < 0 \text{ and } x \geq 0, \\ 0 & \text{if } -1 \leq \hat{\nu} < 0 \text{ and } x < 0. \end{cases} \quad (11)$$

Fig. 15 illustrates CDFs generated from the four gamma fitting methods listed above for a 33-member ensemble containing no dry members, with SEPS values given in parenthesis. A lower SEPS is better so in this example Method (B2) generates the best parametric fit and the dashed green curve would have been selected by the fitting algorithm $\mathcal{F}$.

Finally, a note about goodness of fit and case coverage. The fit of the hybrid gamma distributions were measured using SEPS on a set of over 1 million ACCESS-S 24-hour accumulation (0 UTC aligned) forecast cases from the 33-member ensemble that had at least one wet member. The dataset spanned several seasons across Australia for lead days

1 to 7. The four methods of gamma fitting described in subsection A.1 of this appendix were selected from a pool of eleven methods. Method (B1) was the best performer for fits, as measured by SEPS, but method (E) is needed for edge case coverage when only one ensemble member is wet and method (B2) gives better results on average in situations when the entire ensemble is substantially wet. Combined, methods (E), (B1) and (B2) gave complete case coverage and gave a mean SEPS of 2.348×10^{-3}. Inclusion of method (M) reduced SEPS by 0.6% and inclusion of the remaining seven methods only reduced SEPS by an additional 0.7%. If computational cost is a concern, method (M) could be dropped with slight reduction on average quality of fit.

B Parametrized fits to an ensemble of temperature forecasts

This appendix provides efficient (i.e., vectorizable) methods for fitting ensemble forecasts of temperature to parametric distributions. They are also likely to be suitable for some other forecast variables, such as mean sea level pressure. Parameter estimation methods for four parametric distributions are presented: the 2-parameter normal distribution, the 3-parameter skew normal distribution, the 4-parameter beta distribution and the 4-parameter mixture of two normal distributions with equal variance. Fig. 16 shows these four distributions fit to two sets of daily Tmax forecasts from the 33-member ACCESS-S ensemble. As measured by SEPS, the beta distribution (dotted pink curve) provides the best fit to the data in Fig. 16a while the mixture of two normals (dotted-dash green curve) provides the best fit in Fig. 16b.

In the following, let $\mathbf{x}$ denote a tuple $(x_i)_{i=1}^n$ of ensemble temperature forecasts for one particular forecast case.

B.1 Normal distribution

The CDF F_{n} of the normal distribution with real-valued location parameter μ and positive scale parameter σ is given by

$$F_{\mathrm{n}}(x;\mu,\sigma) = \frac{1}{2}\left(1 + \operatorname{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right)$$

whenever $x \in \mathbb{R}$. Unbiased estimators of μ and σ are given by $\hat{\mu} = \bar{\mathbf{x}}$ and $\hat{\sigma} = \mathrm{sd}(\mathbf{x})$.

B.2 Skew normal distribution

The CDF F_{sn} of the skew normal distribution (Azzalini, 1985, 1986) with real-valued location parameter μ, positive scale parameter σ, and real-valued skew parameter α is given by

$$F_{\mathrm{sn}}(x;\mu,\sigma,\alpha) = F_{\mathrm{n}}\left(\frac{x-\mu}{\sigma};0,1\right) - 2T\left(\frac{x-\mu}{\sigma},\alpha\right)$$

whenever $x \in \mathbb{R}$, where T is Owen's T function.

Parameters are estimated using the method of moments as developed by Arnold et al. (1993) and presented by Lin et al. (2007). Specifically, given the first three sample moments

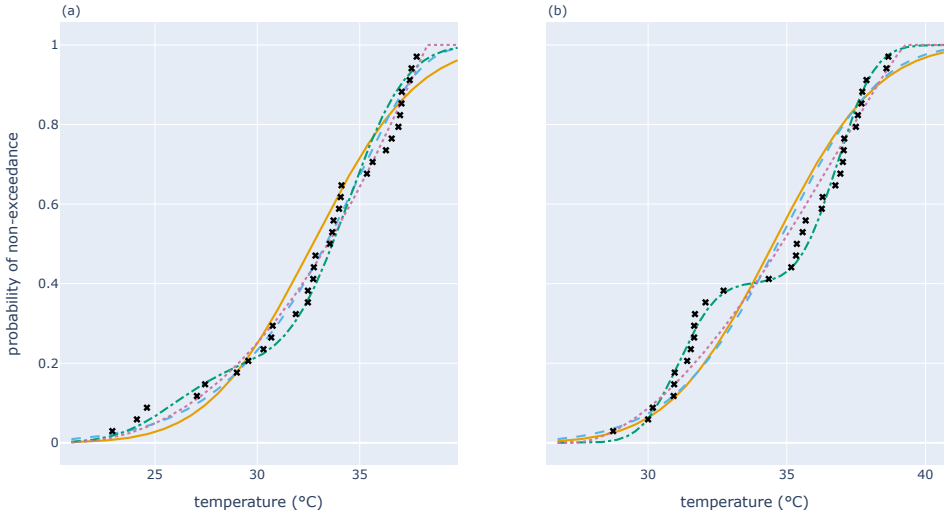

Figure 16: Fitted distributions (curves) to temperature forecasts from the 33-member ACCESS-S ensemble (black crosses). Fitted distributions are selected from the normal (solid gold), skew normal (dashed blue), mixture of two normals with equal variance (dotted-dash green) and beta (dotted pink) families.

$\bar{\mathbf{x}}$, $\mathrm{var}(\mathbf{x})$ and M_3, where

$$M_3 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{\mathbf{x}})^3,$$

we obtain

$$\hat{\mu} = \bar{\mathbf{x}} - a\left(\frac{M_3}{b}\right)^{1/3},$$

$$\hat{\sigma}^2 = \mathrm{var}(\mathbf{x}) + a^2\left(\frac{M_3}{b}\right)^{2/3},$$

$$\hat{\alpha} = \frac{\mathrm{sgn}(M_3)\delta}{\sqrt{1-\delta^2}},$$

where $a = \sqrt{2/\pi}$, $b = (4/\pi - 1)a$ and

$$\delta = \left(a^2 + \mathrm{var}(\mathbf{x})\left(\frac{b}{M_3}\right)^{2/3}\right)^{-1/2}.$$

Note that, depending on the sample, the quantity δ^2 may be greater than or equal to 1, in which case $\hat{\alpha}$ is undefined. The fitting procedure we implemented clips δ so that it lies in the interval $[-15/\sqrt{226}, 15/\sqrt{226}]$. This ensures that $\hat{\alpha}$ is always defined and satisfies $|\hat{\alpha}| \leq 15$. Apart from always providing a skew normal fit, the practical advantage of this implementation is that the magnitude of the skew parameter is never too large, thereby avoiding heavy tails that lead to unrealistic forecasts for the 0.01 or 0.99 quantile levels.

B.3 Mixture of two normal distributions

Often the ensemble can be partitioned into two or more clusters, depending on the synoptic pattern. For example from a 33-member ensemble 10 members may have northerlies over Melbourne, Australia, ahead of a front resulting in higher daytime temperature forecasts, while the remaining 23 members have southerlies in the wake of the front with cooler temperature forecasts. If both clusters are approximately normally distributed, then the entire ensemble will be well modeled as a mixture of two normal distributions.

The CDF F_{mn} of a mixture of two normal distributions is given by

$$F_{\mathrm{mn}}(x; p, \mu_1, \sigma_1, \mu_2, \sigma_2) = pF_{\mathrm{n}}(x; \mu_1, \sigma_1) + (1-p)F_{\mathrm{n}}(x; \mu_2, \sigma_2)$$

whenever $x \in \mathbb{R}$, where without loss of generality, the mixing parameter p satisfies $0 < p \leq 0.5$ and the remaining four parameters μ_1, μ_2, σ_1 and σ_2 are location and scale parameters of their respective normal distributions.

The requirement for computationally fast estimation of these parameters for large gridded multidimensional datasets places restrictions on the estimation procedure employed. The method of moments involves solving a nonic equation for each sample, which is computationally expensive. Maximum likelihood methods are even more prohibitive. Instead, we use the method of moments under the assumption that the two scale parameters are equal. This method, devised by Tan and Chang (1972), and based on earlier work of Rao (1948) and Cohen (1967), requires solving cubic equations.

We summarize the method here, correcting a typographic error of Tan and Chang (1972, Equation 2.5). The method uses the first four k-statistics of the sample $\mathbf{x}$, which are unbiased estimators for the corresponding cumulants and given by

$$\begin{aligned}
\hat{k}_1 &= \bar{\mathbf{x}}, \\
\hat{k}_2 &= \mathrm{var}(\mathbf{x}), \\
\hat{k}_3 &= \frac{n^2}{(n-1)(n-2)} m_3 \quad \text{and} \\
\hat{k}_4 &= \frac{n^2((n+1)m_4 - 3(n-1)m_2^2)}{(n-1)(n-2)(n-3)},
\end{aligned}$$

where the kth central sample moment m_k is given by

$$m_k = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{\mathbf{x}})^k.$$

Define the quantity $\hat{t} = \hat{k}_3^4/\hat{k}_4^3$. Then we find the unique real root z, which satisfies $-1/2 \leq z < 1$, of the cubic equation

$$(4 + 54\hat{t})z^3 - 3z - 1 = 0.$$

This root can be calculated using vectorized methods via Cardano's formula. Then

$$\begin{aligned}
\hat{p} &= 1/2 - \sqrt{3 + 6z}/6 \\
\hat{d} &= (-\hat{k}_4/2)^{1/4} && \text{if } \hat{p} = 1/2 \\
\hat{d} &= \frac{(1-2\hat{p})\hat{k}_4}{2(1-6\hat{p}+6\hat{p}^2)\hat{k}_3} && \text{if } \hat{p} \neq 1/2 \\
\hat{\sigma} &= \left(\hat{k}_2 - 4\hat{d}^2\hat{p}(1-\hat{p})\right)^{1/2} \\
\hat{\mu}_1 &= \hat{k}_1 + 2\hat{d}(1-\hat{p}) \\
\hat{\mu}_2 &= \hat{k}_1 - 2\hat{d}\hat{p}.
\end{aligned}$$

Occasionally this method will fail because the quantity under the square root from which $\hat{\sigma}$ is calculated will be negative. Rejection rates are discussed in subsection B.5 of this appendix.

In practice we clip $\hat{p}$ so that it is at least 0.05, after which the other parameters are estimated. This clip avoids the undesirable situation where one of the normal distributions is being fitted to only one or two data points.

B.4 Beta distribution

The beta distribution provides a flexible family of distributions with finite support. For real-valued location parameter μ, positive scale parameter σ and positive shape parameters α_1 and α_2, the CDF F_{b} of the beta distribution is given by

$$F_{\mathrm{b}}(x; \alpha_1, \alpha_2, \mu, \sigma) = \begin{cases} 0 & \text{if } x < \mu, \\ I\left(\frac{x-\mu}{\sigma}, \alpha_1, \alpha_2\right) & \text{if } \mu \leq x \leq \mu + \sigma, \\ 1 & \text{if } x > \mu + \sigma, \end{cases}$$

where I is the regularized incomplete beta function defined by

$$I(s, a, b) = \Big(\int_0^1 t^{a-1}(1-t)^{b-1}\,\mathrm{d}t \Big)^{-1} \int_0^s t^{a-1}(1-t)^{b-1}\,\mathrm{d}t.$$

This distribution has support on the interval $[\mu, \mu + \sigma]$.

The method of moments can be used to estimate the four parameters of the beta distribution. However, we prefer to use information from the sample tails to estimate the interval $[\mu, \mu + \sigma]$, and then use the method of moments to estimate the two shape parameters. This approach guarantees that the sample $\mathbf{x}$ lies within the estimated interval, which is critical for a weakly-consistent mapping of all ensemble members using the transformation ϕ_{w}.

Given ordered ensemble members $x_1 \leq x_2 \leq \ldots \leq x_n$ with n at least 4, we make the estimate

$$[\hat{\mu}, \hat{\mu} + \hat{\sigma}] = [x_1 - 2(x_4 - x_1)/3, x_n + 2(x_n - x_{n-3})/3].$$

That is, $\hat{\mu} = x_1 - 2(x_4 - x_1)/3$ and $\hat{\sigma} = x_n + 2(x_n - x_{n-3})/3 - \hat{\mu}$. Then defining $u = (\bar{\mathbf{x}} - \hat{\mu})/\hat{\sigma}$ and $v = \mathrm{var}(\mathbf{x})/\hat{\sigma}^2$, the method of moments gives

$$\hat{\alpha}_1 = u\left(\frac{u(1-u)}{v} - 1\right) \quad \text{and} \quad \hat{\alpha}_2 = (1-u)\left(\frac{u(1-u)}{v} - 1\right),$$

provided that $v < u(1-u)$ to ensure that the shape parameter estimates are positive.

B.5 Assessment of methods for temperature fits

All four methods for obtaining parametric fits to ensemble temperature forecasts were evaluated using the 33-member ACCESS-S ensemble. The dataset included 678040 forecast cases for daily Tmin at AWS locations across Australia spanning the period 4 November 2022 to 7 May 2023, for lead days 1 to 7.

Based on our implementation of these methods and dataset size, parameter estimation for the skew normal and beta distributions were about six times slower to compute than that for the normal distribution while for the mixture of two normals it about 60 times greater, though this could be reduced using dask parallel processing. Fits for the mixture of two normals were calculated in 84% of cases, with the remaining cases rejected due to negative arguments in the square root of the formula for $\hat{\sigma}$.

When all four methods were used the mean error ϵ, as measured by SEPS, was 1.84×10^{-3}. If only the beta and mixture of two normals fits were used then the resulting mean error was 1.08ϵ. The best single fitting method was the beta fits, with a mean error of 1.37ϵ, followed by skew normal (1.53ϵ) and normal (1.94ϵ). The mixture of two normals cannot be used to fit all the forecasts alone, but performed marginally worse than the beta distribution in those cases for which fits were produced. Results for daily Tmax using a similar dataset were comparable.

C Reconstruction of the ADFD quantile function

This appendix outlines the process for reconstructing the ADFD quantile function F^{-1} using the PoE and quantile ADFD forecasts from Table 1. Note that the domain of the quantile function F^{-1} is the closed unit interval $[0, 1]$.

1. The ADFD PoP forecast p is interpreted as the probability of at least 0.2 mm. That is, $F^{-1}(1-p) = 0.2$. This interpretation is consistent with some internal Bureau processes, and may be justified by the fact that rainfall measurements from AWSs have a resolution of 0.2 mm.

2. Set $F^{-1}(\alpha) = 0$ whenever $0 \leq \alpha \leq 0.99 - p$, which covers all percentile levels associated with no rainfall.

3. Set $F^{-1}(1-p_\theta) = \theta$ for each PoE forecast p_θ associated with thresholds $\theta = 1, 5, 10, 15, 25, 25$.

4. Where not already set, set $F^{-1}(\alpha) = q_\alpha$ for each of the four quantile forecasts q_α with levels $\alpha = 0.25, 0.5, 0.75, 0.9$.

5. Using a Weibull distribution tail, extrapolate quantiles for levels 0.95, 0.96, 0.97, 0.98, 0.99. Only keep extrapolated values whose quantile levels are higher than $1 - \mathbb{P}(Y > 50)$ and whose extrapolated quantiles exceed 50mm. In most cases, these extrapolated quantiles are discarded because $\mathbb{P}(Y > 50) \leq 0.01$. Full details of this extrapolation step are given after this list.

6. All the values found so far are percentile values. Fill the remaining percentiles using linear interpolation.

Table 4: Illustrative example of the information available about the tail of the distribution of an ADFD daily precipitation forecast. This was not sampled from an actual ADFD forecast.

ADFD forecast type	quantile value (mm)	quantile level
probability of exceeding (PoE) 25 mm	25	0.70
probability of exceeding (PoE) 50 mm	50	0.87
75th percentile	40	0.75
90th percentile	51	0.90

7. Reorder percentile values (if necessary) so the they are non-decreasing as a function of percentile level. This step is occasionally needed due to probabilistic inconsistencies between different ADFD forecasts, caused by rounding or by manual editing where consistency checks have not been applied.

8. Fill remaining values of F^{-1} via linear interpolation and extrapolation.

The extrapolation step assumes that the tail of the distribution has the tail of a Weibull distribution. The Weibull distribution is embedded within the larger family of generalized gamma distributions. The reason for using it is twofold: (1) the mathematics is tractable for large gridded datasets and (2) the Weibull distribution is sometimes used in the production of ADFD forecasts. In what follows we describe a number of important steps to ensure that extrapolated quantiles are not unreasonably extreme.

The predictive distribution of precipitation Y could be modeled using a CDF G given by

$$G(x) = \mathbb{P}(Y \leq x) = 1 - p \exp\left(-\left(\frac{x}{\sigma}\right)^{\xi}\right) \tag{12}$$

where $x \geq 0$, p is the PoP, σ is a positive scale parameter and ξ is a positive shape parameter. Note that the conditional distribution of precipitation given that precipitation occurs is a Weibull distribution. The graphs of the CDFs of three such distributions, each with $p = 0.8$, are illustrated in Fig. 17.

We begin by illustrating the extrapolation step with a concrete example and then end with formulae to handle the general case. Suppose that the probability p of precipitation is 0.8 and that the known upper part of the ADFD distribution is given by Table 4. The two forecasts with the highest quantile levels (0.87 and 0.9) could be used in conjunction with p to fit a distribution G. This distribution is shown as a dotted gray curve in Fig. 17. It passes through through the black circle (representing p), the upper solid dark orange circle (representing the 0.87-quantile) and the upper solid blue square (representing the 0.9-quantile). However, we prefer not to do this, since ADFD PoE forecasts are rounded, whereas ADFD percentile forecasts are not, and when a PoE and percentile forecast are close on the CDF, rounding can have a large impact on extrapolated values.

Instead, we obtain two fitted curves: one fitting the two PoE forecasts (solid dark orange circles) and the other fitting the two percentile forecasts (solid blue squares). The fitted distributions are shown as gold and light blue curves. The five tail percentiles (open

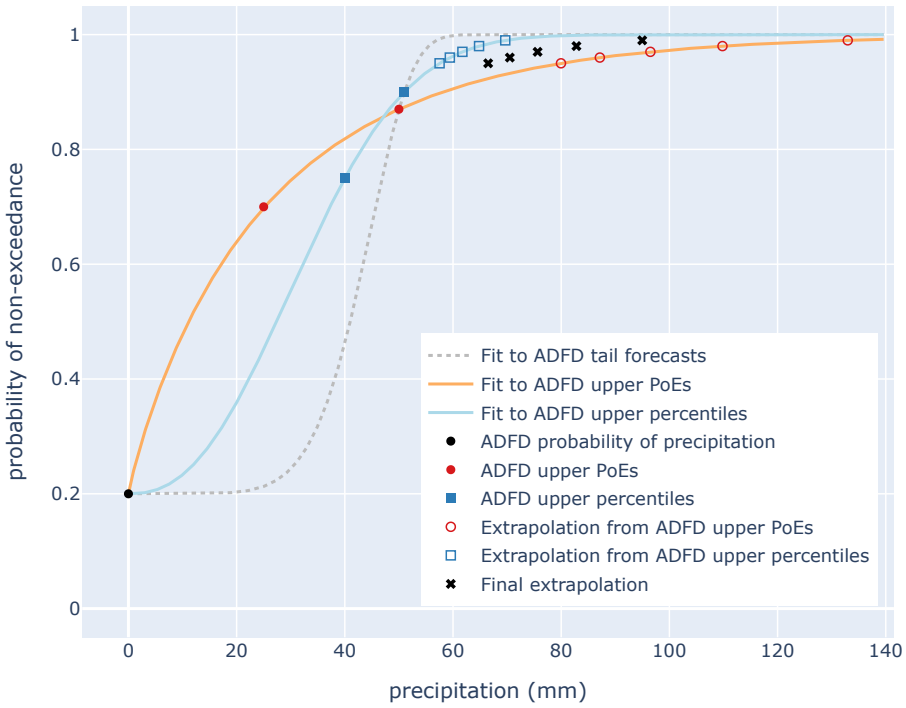

Figure 17: Illustrative example of the Weibull extrapolation method.

dark orange circles and open blue squares) are extrapolated using the tails of these curves. This gives two sets of possible upper percentiles. The final extrapolated tail percentiles (black crosses) are a weighted average of these two sets. The weight is determined by how high the quantile levels are in the ADFD inputs. In this case, the dark orange extrapolated percentiles get a weight of 0.4 while the blue extrapolated percentiles get a weight of 0.6, because the quantile levels of the closed dark orange circles are slightly lower overall relative to the quantile levels of the closed blue squares.

Mathematical details of this extrapolation algorithm are as follows. Suppose that the PoP is p and that two quantiles q_1 and q_2 at quantile levels α_1 and α_2 are known. We wish to find parameters σ and ξ of the fitting distribution G. The quantiles yield the equations $G(q_1) = \alpha_1$ and $G(q_2) = \alpha_2$, whereby we obtain two equations to solve for σ and ξ. Their solutions are

$$\xi = \frac{\log(c_2/c_1)}{\log(q_2/q_1)},$$
$$\sigma = \frac{q_2}{\exp(\xi^{-1}\log(c_2))},$$

where

$$c_1 = -\log\left(1-(\alpha_1-1+p)/p\right),$$
$$c_2 = -\log\left(1-(\alpha_2-1+p)/p\right).$$

In practice, if $\xi < 0.9$ we change its value to 0.9 prior to solving for σ. This avoids unrealistically heavy extrapolated tails, noting that if $\xi = 1$ then the tail is that of an exponential distribution. Once values for ξ and σ are obtained, tail percentiles can be calculated from Equation (12) by taking (say) $G(x) = 0.99$ and solving for x to obtain the 99th percentile.

To determine weights, define the weighting function W, which is a continuous piecewise linear function of the quantile level α, by

$$W(\alpha) = \begin{cases} 0, & 0 < \alpha < 0.75, \\ (\alpha - 0.75)/0.15, & 0.75 \le \alpha \le 0.9, \\ 1, & 0.9 < \alpha < 1. \end{cases}$$

The weight w placed on percentiles extrapolated from the two ADFD PoE forecasts (i.e., the solid dark orange circles of Fig. 17) is given by $w = (W(\alpha_1)+W(\alpha_2))/2$, where α_1 and α_1 are the quantile levels associated with those two forecasts. A weight of $1-w$ is placed on percentiles extrapolated from the two percentile forecasts (i.e., the solid blue squares in Fig. 17). In Table 4 and Fig. 17, $w = 0.4$. However, were the two solid dark orange circles both to the left of the two solid blue squares, the method would give $w = 0$. On the other hand, were the two solid dark orange circles both to the right of the two solid blue squares, then the method would give $w = 1$.

References

Arnold, B. C., R. J. Beaver, R. A. Groeneveld, and W. Q. Meeker, 1993: The nontruncated marginal of a truncated bivariate normal distribution. *Psychometrika*, **58**, 471–488, doi:10.1007/BF02294652.

Azzalini, A., 1985: A class of distributions which includes the normal ones. *Scandinavian journal of statistics*, 171–178.

Azzalini, A., 1986: Further results on a class of distributions which includes the normal ones. *Statistica*, **46 (2)**, 199–208, doi:10.6092/issn.1973-2201/10421.

Balakrishnan, N., and J. Wang, 2000: Simple efficient estimation for the three-parameter gamma distribution. *Journal of Statistical Planning and Inference*, **85 (1-2)**, 115–126, doi:10.1016/S0378-3758(99)00074-9.

Brown, A., S. Milton, M. Cullen, B. Golding, J. Mitchell, and A. Shelly, 2012: Unified modeling and prediction of weather and climate: A 25-year journey. *Bulletin of the American Meteorological Society*, **93 (12)**, 1865–1877, doi:10.1175/BAMS-D-12-00018.1.

Brunet, G., S. Jones, P. M. Ruti, and Coauthors, 2015: *Seamless prediction of the Earth System: from minutes to months*. World Meteorological Organization.

Bureau of Meteorology, 2018: *Upgrades to the Operational PME System*, Vol. 116. BNOC Operations Bulletin, URL https://bom.gov.au/australia/charts/bulletins/apob116_external.pdf.

Bureau of Meteorology, 2023: *User guide: Australian Digital Forecast Database (ADFD)*. Bureau of Meteorology, URL http://www.bom.gov.au/catalogue/adfdUserGuide.pdf.

Burgeno, J. N., and S. L. Joslyn, 2020: The impact of weather forecast inconsistency on user trust. *Weather, climate, and society*, **12 (4)**, 679–694, doi:10.1175/WCAS-D-19-0074.1.

Cohen, A. C., 1967: Estimation in mixtures of two normal distributions. *Technometrics*, **9 (1)**, 15–28.

de Burgh-Day, C., and F. Dillon, 2021: *A hybrid parametrisation for precipitation probability of exceedance data*, Vol. 52. Bureau of Meteorology Research Report, URL http://www.bom.gov.au/research/publications/researchreports/BRR-052.pdf.

de Burgh-Day, C., and R. Taggart, 2023: *A computationally efficient method for fitting smooth parametrized curves to precipitation distributions*, Vol. 81. Bureau of Meteorology Research Report, URL http://www.bom.gov.au/research/publications/researchreports/BRR-081.pdf.

Diebold, F. X., and R. S. Mariano, 2002: Comparing predictive accuracy. *Journal of Business & economic statistics*, **20 (1)**, 134–144, doi:10.1198/073500102753410444.

Ferro, C. A., D. S. Richardson, and A. P. Weigel, 2008: On the effect of ensemble size on the discrete and continuous ranked probability scores. *Meteorological Applications: A journal of forecasting, practical applications, training techniques and modelling*, **15 (1)**, 19–24, doi:10.1002/met.45.

Gneiting, T., 2011: Making and evaluating point forecasts. *Journal of the American Statistical Association*, **106 (494)**, 746–762, doi:10.1198/jasa.2011.r10138.

Gneiting, T., and A. E. Raftery, 2007: Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, **102 (477)**, 359–378, doi:10.1198/016214506000001437.

Griffiths, D., and A. Jayawardena, 2022: *AutoFcst: A Coherent Set of Forecast Grids*, Vol. 71. Bureau of Meteorology Research Report, URL http://www.bom.gov.au/research/publications/researchreports/BRR-071.pdf.

Griffiths, M., P. Smith, H. Yan, C. Spillman, and G. Young, 2023: *ACCESS-S2: Updates and improvements to postprocessing pipeline*, Vol. 82. Bureau of Meteorology Research Report, URL http://www.bom.gov.au/research/publications/researchreports/BRR-082.pdf.

Hering, A. S., and M. G. Genton, 2011: Comparing spatial predictions. *Technometrics*, **53 (4)**, 414–425, doi:10.1198/TECH.2011.10136.

Hosking, J. R. M., J. R. Wallis, and E. F. Wood, 1985: Estimation of the generalized extreme-value distribution by the method of probability-weighted moments. *Technometrics*, **27 (3)**, 251–261, doi:10.1080/00401706.1985.10488049.

Hoskins, B., 2013: The potential for skill across the range of the seamless weather-climate prediction problem: a stimulus for our science. *Quarterly Journal of the Royal Meteorological Society*, **139 (672)**, 573–584, doi:10.1002/qj.1991.

Hudson, D., and Coauthors, 2017: ACCESS-S1 the new Bureau of Meteorology multi-week to seasonal prediction system. *Journal of Southern Hemisphere Earth Systems Science*, **67 (3)**, 132–159, doi:10.1071/ES17009.

Jones, D. A., W. Wang, and R. Fawcett, 2009: High-quality spatial climate datasets for australia. *Australian Meteorological and Oceanographic Journal*, **58 (4)**, 233, doi:10.22499/2.5804.003.

Just, A., and M. Foley, 2020: Streamlining the graphical forecast process. *Journal of Southern Hemisphere Earth Systems Science*, **70 (1)**, 108–113, doi:10.1071/ES19047.

Kober, K., G. C. Craig, C. Keil, and A. Dörnbrack, 2012: Blending a probabilistic nowcasting method with a high-resolution numerical weather prediction ensemble for convective precipitation forecasts. *Quarterly Journal of the Royal Meteorological Society*, **138 (664)**, 755–768, doi:10.1002/qj.939.

Kumar, A., and R. Murtugudde, 2013: Predictability, uncertainty and decision making: A unified perspective to build a bridge from weather to climate. *Current Opinion in Environmental Sustainability*, **5 (3-4)**, 327–333, doi:10.1016/j.cosust.2013.05.009.

Lin, T. I., J. C. Lee, and S. Y. Yen, 2007: Finite mixture modelling using the skew normal distribution. *Statistica Sinica*, 909–927.

Loveday, N., and Coauthors, 2024: The jive verification system and its transformative impact on weather forecasting operations. *Bulletin of the American Meteorological Society*, **105 (11)**, E2047–E2063, doi:10.1175/BAMS-D-23-0267.1.

Mase, A. S., and L. S. Prokopy, 2014: Unrealized potential: A review of perceptions and use of weather and climate information in agricultural decision making. *Weather, Climate, and Society*, **6 (1)**, 47–61, doi:10.1175/WCAS-D-12-00062.1.

Matheson, J. E., and R. L. Winkler, 1976: Scoring rules for continuous probability distributions. *Management science*, **22 (10)**, 1087–1096, doi:10.1287/mnsc.22.10.1087.

National Weather Service, 2024: *Seasonal Outlooks: Probability of Exceedence (POE) Maps*. National Weather Service, URL https://www.cpc.ncep.noaa.gov/products/predictions/long_range/poe_index.php?lead=1&var=p.

Rao, C. R., 1948: The utilization of multiple measurements in problems of biological classification. *Journal of the Royal Statistical Society. Series B (Methodological)*, **10 (2)**, 159–203.

Ren, H.-L., and Coauthors, 2023: Seamless prediction in china: A review. *Advances in Atmospheric Sciences*, **40 (8)**, 1501–1520, doi:10.1007/s00376-023-2335-z.

Ruti, P. M., and Coauthors, 2020: Advancing research for seamless earth system prediction. *Bulletin of the American Meteorological Society*, **101 (1)**, E23–E35, doi:10.1175/BAMS-D-17-0302.1.

Saunders, T., 2023: *Forecast rain a welcome relief for farmers after bureau's grim June outlook*. Australian Broadcasting Corporation, URL https://www.abc.net.au/news/2023-06-03/forecast-rain-relief-for-farmers-after-dry-outlook-for-june/102435690.

Scheufele, K., K. Kober, G. C. Craig, and C. Keil, 2014: Combining probabilistic precipitation forecasts from a nowcasting technique with a time-lagged ensemble. *Meteorological Applications*, **21 (2)**, 230–240, doi:10.1002/met.1381.

Senior, C., and Coauthors, 2008: Seamless prediction. *MOSAC Paper*, **13**.

Shukla, J., 2009: Seamless prediction of weather and climate: A new paradigm for modeling and prediction research. *Climate Test Bed Joint Seminar Series*, Citeseer, 8.

Sloughter, J. M. L., A. E. Raftery, T. Gneiting, and C. Fraley, 2007: Probabilistic quantitative precipitation forecasting using bayesian model averaging. *Monthly weather review*, **135 (9)**, 3209–3220, doi:10.1175/MWR3441.1.

Taggart, R., N. Loveday, and D. Griffiths, 2022: A scoring framework for tiered warnings and multicategorical forecasts based on fixed risk measures. *Quarterly Journal of the Royal Meteorological Society*, **148 (744)**, 1389–1406, doi:10.1002/qj.4266.

Tan, W., and W. Chang, 1972: Some comparisons of the method of moments and the method of maximum likelihood in estimating parameters of a mixture of two normal densities. *Journal of the American Statistical Association*, **67 (339)**, 702–708, doi:10.1080/01621459.1972.10481282.

Trotta, B., and Coauthors, 2024: RainForests: a machine-learning approach to calibrating NWP precipitation forecasts. *Weather and forecasting*, doi:10.1175/WAF-D-23-0211.1.

Vitart, F., and Coauthors, 2008: The new VarEPS-monthly forecasting system: A first step towards seamless prediction. *Quarterly Journal of the Royal Meteorological Society*, **134 (636)**, 1789–1799, doi:10.1002/qj.322.

Wedd, R., and Coauthors, 2022: ACCESS-S2: the upgraded Bureau of Meteorology multi-week to seasonal prediction system. *Journal of Southern Hemisphere Earth Systems Science*, **72 (3)**, 218–242, doi:10.1071/ES22026.