

Calibrating Hourly Precipitation Forecasts with Daily Observations

C. CATTOËN

National Institute of Water and Atmospheric Research, Christchurch, New Zealand

D. E. ROBERTSON

Commonwealth Scientific and Industrial Research Organisation, Melbourne, Victoria, Australia

J. C. BENNETT

Commonwealth Scientific and Industrial Research Organisation, Melbourne, Victoria, and Institute for Marine and Antarctic Studies, University of Tasmania, Hobart, Australia

Q. J. WANG

University of Melbourne, Melbourne, Victoria, Australia

T. K. CAREY-SMITH

National Institute of Water and Atmospheric Research, Wellington, New Zealand


(Manuscript received 16 October 2019, in final form 26 February 2020)

ABSTRACT

Calibrated high-temporal-resolution precipitation forecasts are desirable for a range of applications, for example, flood prediction in fast-rising rivers. However, high-temporal-resolution precipitation observations may not be available to support the establishment of calibration methods, particularly in regions with low population density or in developing countries. We present a new method to produce calibrated hourly precipitation ensemble forecasts from daily observations. Precipitation forecasts are taken from a high-resolution convective-scale numerical weather prediction (NWP) model run at the hourly time step. We conduct three experiments to develop the new calibration method: (i) calibrate daily precipitation totals and disaggregate daily forecasts to hourly; (ii) generate pseudohourly observations from daily precipitation observations, and use these to calibrate hourly precipitation forecasts; and (iii) combine aspects of (i) and (ii). In all experiments, we use the existing Bayesian joint probability model to calibrate the forecasts and the well-known Schaake shuffle technique to instill realistic spatial and temporal correlations in the ensembles. As hourly observations are not available, we use hourly patterns from the NWP as the template for the Schaake shuffle. The daily member matching method (DMM), method (iii), produces the best-performing ensemble precipitation forecasts over a range of metrics for forecast accuracy, bias, and reliability. The DMM method performs very similarly to the ideal case where hourly observations are available to calibrate forecasts. Overall, valuable spatial and temporal information from the forecast can be extracted for calibration with daily data, with a slight trade-off between forecast bias and reliability.

1. Introduction

Two trends have emerged in the development of new streamflow forecasting systems: (i) a shift from deterministic to ensemble streamflow predictions (Alfieri et al. 2013; Cloke and Pappenberger 2009; Demargne et al. 2014; Thielen et al. 2009), and (ii) a move toward national/continental scale systems that attempt to describe hydrological fluxes for all reaches over a given domain (Adams and Pagano 2016; Bell et al. 2017; Emerton et al. 2016; Maxey et al. 2012). Meeting these twin

 Denotes content that is immediately available upon publication as open access.

 Supplemental information related to this paper is available at the Journals Online website: <https://doi.org/10.1175/JHM-D-19-0246.s1>.

Corresponding author: C. Cattoën, celine.cattoen-gilbert@niwa.co.nz

DOI: 10.1175/JHM-D-19-0246.1

© 2020 American Meteorological Society. For information regarding reuse of this content and general copyright information, consult the [AMS Copyright Policy](https://www.ametsoc.org/PUBSReuseLicenses) (www.ametsoc.org/PUBSReuseLicenses).

aims offers clear benefits to forecast users: ensemble forecasts are usually more accurate than deterministic predictions and give an explicit estimate of forecast uncertainty (Gneiting and Katzfuss 2014; Gneiting et al. 2007), while extensive spatial coverage gives forecast information in reaches/basins where it was previously unavailable.

New Zealand's National Institute of Water and Atmospheric Research (NIWA) is developing a national scale ensemble streamflow forecasting system for New Zealand, with the aim of informing water management and emergency agencies. New Zealand's mountainous topography leads to precipitation that varies sharply in space and time, and in turn to rivers that can rise very quickly (Cattoën et al. 2016; Woods et al. 2006). These catchment characteristics provide similar challenges present in many mountainous regions (Rossa et al. 2011). To produce useful streamflow forecasts in New Zealand—particularly flood forecasts—requires (i) high-resolution precipitation forecasts to account for orographic effects due to steep mountains and (ii) hydrological models run at an hourly time step.

Rainfall forecasts are produced by NIWA at very high spatial resolution (1.5 km); however, the computational cost of this numerical weather prediction (NWP) model means that it can only produce deterministic forecasts. Fortunately, methods are available to produce ensemble precipitation forecasts through statistical calibration of deterministic NWP outputs (see review by Li et al. 2017). Statistical calibration offers the additional benefits of correcting biases and ensuring “coherence”—i.e., ensuring forecasts are at least as accurate as climatology forecasts (Zhao et al. 2017). These properties are essential prerequisites for using forecasts to force hydrological models.

A key requirement of statistical calibration is the availability of observations at the time step of interest—in our case, hourly precipitation. New Zealand has a sparse rainfall gauge network in relation to the very high spatial variability of rainfall in mountainous regions, meaning that hourly rainfall observations are not available at the national scale. There is a rain radar network covering much of New Zealand, but it is only available on a commercial basis so was not used in this study. Additionally, due to the complex terrain of many New Zealand catchments, radar accuracy can be degraded, and coverage significantly limited. However, daily precipitation data are available across New Zealand in the form of the interpolated and mass-corrected Virtual Climate Station Network (VCSN) dataset (Tait et al. 2006). The VCSN interpolates observed meteorological values onto a grid covering New Zealand at a 5-km spatial resolution at a daily time step. The mass correction is necessary to overcome underestimation of precipitation in mountainous regions (Andréassian et al. 2010; Bartolini et al. 2011; Beck et al. 2019; Hamon 1973; Valéry et al.

2010). The mass correction is performed by comparing rainfall and long-term streamflow records and correcting rainfall to ensure mass balance (Woods et al. 2006).

Lack of subdaily rainfall observations is a problem facing many regions where calibrated rainfall forecasts could be useful. Hourly precipitation datasets with extensive national or continental coverage are unusual—particularly in developing nations (Gruber and Levizzani 2008)—whereas daily precipitation datasets are more common and available over large domains [e.g., for the United States (Peterson et al. 1997), Canada (Vincent and Mekis 2006), and Australia (Jones et al. 2009) as well as global datasets (Beck et al. 2019)].

In this study, we aim to establish a new method to calibrate hourly precipitation forecasts from daily observations. At the time of writing, we are unaware of existing work addressing this issue. We base this on an existing calibration method (Robertson et al. 2013; Shrestha et al. 2015), which combines a Bayesian joint probability (BJP) model to calibrate forecasts with the Schaake shuffle (Clark et al. 2004) to order calibrated forecast ensemble members in space and time. We note that other reordering methods are available, notably ensemble copula coupling (ECC) (Schefzik et al. 2013). ECC is attractive in cases of data paucity because it differs from the Schaake shuffle in its choice of dependency template. For the Schaake shuffle, the template is chosen from past observations, whereas for ECC, the template is the uncalibrated ensemble forecast. For this paper, the uncalibrated forecast is deterministic and thus ECC could not be used.

We conduct three experiments: (i) calibrating to daily observations, and then disaggregating calibrated daily forecasts to hourly; (ii) synthesizing hourly observations from daily data using temporal and spatial patterns from the NWP, and then calibrating directly to these “pseudohourly” observations; and (iii) calibrating to both daily observations and to pseudohourly observations, and using calibrated daily rainfall forecasts to correct daily totals of pseudohourly calibrated forecasts. We compare these to the ideal case where hourly observations are available and demonstrate that the third experiment produces the best performing ensemble forecasts.

The paper is structured as follows: section 2 describes the catchment, observations, and NWP predictions used in this study. Section 3 describes the implementation of the BJP modeling approach for postprocessing subdaily rainfall predictions (control), three experiments to postprocess hourly forecasts with daily data, and the methods used to verify forecasts. Section 4 presents the results of forecast verification and compares the three experiments to the control obtained with hourly data. Section 5 discusses the potential limitations of the methods presented and identifies possible extensions. Section 6 summarizes and concludes the paper.

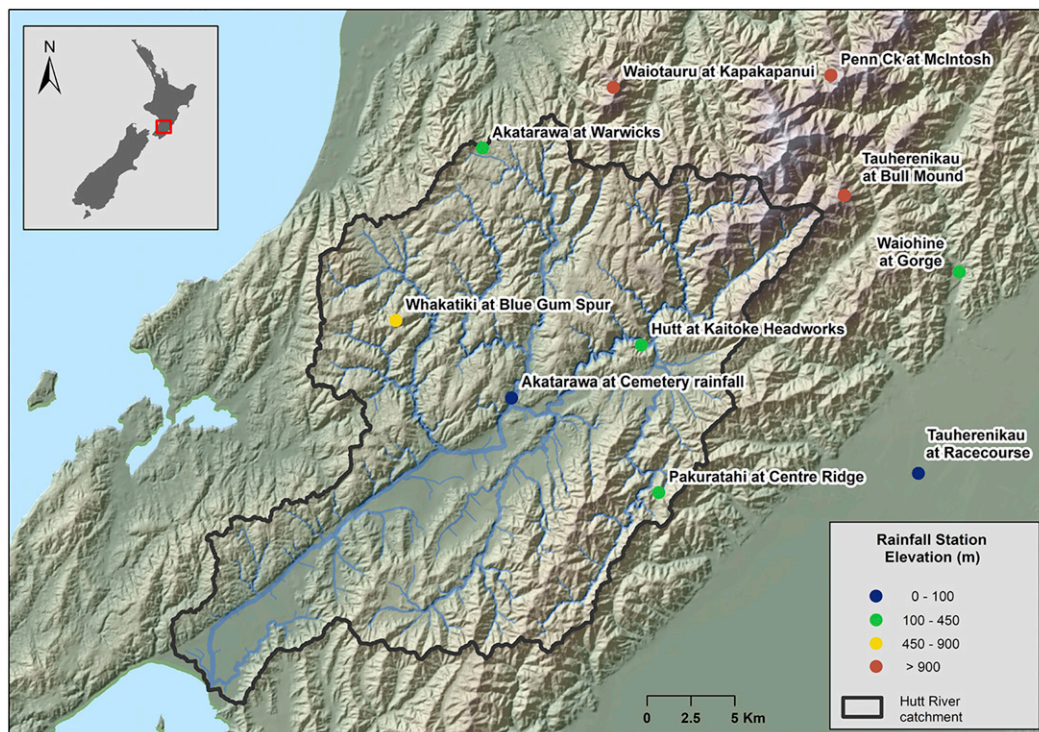


FIG. 1. Map of the Hutt catchment in the Wellington region in New Zealand, with location of rainfall stations and topography.

2. Catchment, data, and NWP model

a. Catchment and data

We use the Hutt catchment in the Wellington region of New Zealand to test our method (Fig. 1). The elevation range of the catchment is large, with mountainous areas (the Tararua and Rimutaka ranges) in the northwest, and an extensive floodplain in the lower reaches. The catchment features a very steep precipitation gradient, with annual rainfall ranging from <900 to >5000 mm over its area of 558 km^2 (Ballinger et al. 2011; Wellington Regional Council 1995) (Table 1).

While our method will eventually be deployed with the VCSN dataset, to test it thoroughly we require hourly observations as a benchmark. We therefore use gauged hourly precipitation for this study. The catchment is densely gauged: hourly observed precipitation data are available from 10 automatic meteorological stations with tipping-bucket rain gauges (Table 1). The tipping buckets are 0.5 mm in volume. For most of the stations an historical archive of hourly precipitation is available since 1972; however, we restrict the records to a 3-yr period, 2014–17, to match the precipitation forecast archive (section 2b). These three years include a moderately dry year (2014/15), a near-normal year (2016/17), and one with rainfall in the top 5% (2015/16),

based on the past 40 years of records. Two out of the 10 rainfall stations have missing data ranging from 0.6% to 10.7% (Table 1).

As already noted, we wish to test our methods on daily observations in order to be compatible with the VCSN dataset. To produce daily observations from the precipitation gauging stations, we sum hourly data for each 24-h period beginning at 2100 UTC, the same 24-h aggregation period for which the VCSN observations are calculated.

b. NWP model

Rainfall forecasts are generated by the New Zealand Convective Scale Model (NZCSM), a local implementation of the U.K. Met Office Unified Model System (UM), which has been run operationally since 2014. NZCSM is run as a deterministic model, with a grid resolution of 1.5 km and outputs archived at a 30-min time step. Forecasts are run to a lead time of 36 h. NZCSM takes its forcing from the New Zealand Limited Area Model (NZLAM), a regional NWP model run at a 12-km resolution that uses lateral boundary conditions from the global version of the UM run by the Met Office. NZCSM's initial conditions are generated via a pseudo data assimilation scheme that optimally combines the large-scale features of the NZLAM forecast. The 1.5-km

TABLE 1. Rainfall station information for the case study catchment during the 3-yr period May 2014–17.

Gauge name	Gauge No.	Elevation (m)	Mean precipitation (mm yr ⁻¹)	Missing data (%)	Latitude (°S)	Longitude (°E)
Tauherenikau at Racecourse	15132710	40	837	0	41.12	175.38
Akatarawa at Cemetery rainfall	150108	100	1858	0	41.09	175.09
Waiohine at Gorge	1503191	140	2186	10.7	41.01	175.376
Hutt at Kaitoke Headworks	150201	190	2254	0	41.06	175.19
Whakatiki at Blue Gum Spur	150010	335	2313	0	41.05	175.02
Akatarawa at Warwicks	59007	345	2775	0.6	40.96	175.08
Pakuratahi at Centre Ridge	151202	510	2017	0	41.13	175.20
Tauherenikau at Bull Mound	59310	1030	4452	0	40.98	175.32
Waiotauru at Kapakapanui	59104	1102	3115	0.6	40.92	175.17
Penn Ck at McIntosh	59201	1286	5973	0	40.91	175.31

grid resolution of the NZCSM allows an accurate representation of the New Zealand topography, which is especially beneficial in mountainous regions. NZCSM forecasts are issued four times a day, at 0300, 0900, 1500, and 2100 UTC. To avoid ambiguity, we refer to each forecast issue time as a *cycle*, and define the cycles as 0300 cycle, 0900 cycle, 1500 cycle, and 2100 cycle.

An archive of real-time predictions for the ~3-yr period from 1 May 2014 to 31 May 2017 is available for this study (approximately 4500 forecasts, or ~1157 forecasts for each cycle). While a longer record is desirable, it is unavailable due to significant model upgrades to the NZCSM model in 2017, which substantially improved the outputs of the model.

To minimize potentially undesirable model spinup effects, we avoid the use of the first 6 h of the forecast in the calibration process, which include the forecast incremental analysis and pseudodata assimilation time period (Cattoën et al. 2016). Analysis of rainfall properties as a function of lead time shows that while most of the spinup effects are resolved within the first 3 h, there can still be some impact after 5 to 6 h (Cattoën et al. 2019). The 6-h offset was chosen also to better match the available daily observation period (0900–0900 local time).

A known problem with NZCSM predictions (and some other UM implementations; e.g., Stratton et al. 2018) is occasional predictions of unrealistically large rainfalls. The excess rain is caused by the model failing to conserve mass in certain circumstances. To remove unrealistic values, we fit a log-sinh (Wang et al. 2012) transformed normal distribution to the hourly forecasts for each combination of station, lead time, and issue cycle. We then compute the hourly forecast value with an exceedance probability of 1 in 100 years according to the fitted distribution. We refer to this value as p_{extreme} . Any forecast values greater than p_{extreme} are set to p_{extreme} . These adjustments are made to at most five forecast values (of ~1157) for

each lead time. This method for dealing with unrealistically large rainfalls was used in Wang et al. (2019a,b).

3. Methods

a. Forecast calibration

Our forecast calibration uses a censored bivariate normal distribution to relate transformed NWP rainfall forecasts to transformed observations (Robertson et al. 2013). The transformation, censored bivariate normal distribution, and the parameter estimation procedure are described in section 3a(1), section 3a(2), and appendix A, respectively. The calibrated forecasts are reordered with the Schaake shuffle [section 3a(4)], to ensure realistic temporal and spatial rank structures in the ensemble. This postprocessing method has been applied extensively in Australia (Shrestha et al. 2015), where it forms the basis of a preoperational ensemble streamflow forecasting system (Bennett et al. 2014).

1) DATA CENSORING AND TRANSFORMATION

Deterministic rainfall forecasts x are rounded to the nearest 0.01 mm and a censor threshold of 0 mm is used for the parameter estimation procedure. For observations y , a censor threshold of 0.5 mm is applied, to be consistent with the tipping buckets used in this study of 0.5 mm in volume.

We first scale rainfall forecasts x and observations y :

$$\begin{aligned} x' &= \frac{5x}{x_{\max}}, \\ y' &= \frac{5y}{y_{\max}}, \end{aligned} \quad (1)$$

where x_{\max} and y_{\max} are the maximum values of x and y over the full data period used.

The scaling forces the transformed forecasts and observations, \tilde{x} and \tilde{y} , to take a common range of values. This simplifies the application of a prior in parameter estimation, as described in [appendix A](#). The scaling is also of practical value when applying the calibration over a wide range of catchments, as it means each model parameter will take a common range of values for all catchments.

After scaling, the data are transformed with the log-sinh transformation ([Wang et al. 2012](#)) to normalize x' and y' , and homogenize their variances:

$$\begin{aligned}\tilde{x} &= \frac{1}{b_x} \log[\sinh(a_x + b_x x')], \\ \tilde{y} &= \frac{1}{b_y} \log[\sinh(a_y + b_y y')],\end{aligned}\quad (2)$$

where a_x and b_x are transformation parameters for x' , and a_y and b_y for y' .

2) BIVARIATE NORMAL DISTRIBUTION

We assume \tilde{x} and \tilde{y} follow a bivariate normal distribution

$$p(\tilde{x}, \tilde{y}) \sim N(\mu, \Sigma), \quad (3)$$

where

$$\mu = \begin{bmatrix} \mu_{\tilde{x}} \\ \mu_{\tilde{y}} \end{bmatrix}, \quad (4)$$

and $\mu_{\tilde{x}}$ is the mean of \tilde{x} and $\mu_{\tilde{y}}$ is the mean of \tilde{y} ;

$$\Sigma = \begin{bmatrix} \sigma_{\tilde{x}}^2 & \rho_{\tilde{x}\tilde{y}} \sigma_{\tilde{x}} \sigma_{\tilde{y}} \\ \rho_{\tilde{x}\tilde{y}} \sigma_{\tilde{x}} \sigma_{\tilde{y}} & \sigma_{\tilde{y}}^2 \end{bmatrix}, \quad (5)$$

where $\sigma_{\tilde{x}}$ and $\sigma_{\tilde{y}}$ are the standard deviations of \tilde{x} and \tilde{y} , respectively, and $\rho_{\tilde{x}\tilde{y}}$ is a correlation coefficient.

A set of parameters $\theta = [a_x \ b_x \ \mu_{\tilde{x}} \ \sigma_{\tilde{x}} \ a_y \ b_y \ \mu_{\tilde{y}} \ \sigma_{\tilde{y}} \ \rho_{\tilde{x}\tilde{y}}]$ is inferred for each rain gauge and for each lead time. Parameters are inferred with maximum a posteriori (MAP) estimation with zeros treated as censored data as detailed in [appendix A](#).

3) GENERATING A CALIBRATED FORECAST

Given a parameter set θ , we can define a univariate normal distribution:

$$f(\tilde{y}|\tilde{x}, \theta) \sim N(\mu_{\tilde{y}|\tilde{x}, \theta}, \sigma_{\tilde{y}|\tilde{x}, \theta}^2) \quad (6)$$

with mean

$$\mu_{\tilde{y}|\tilde{x}, \theta} = \mu_{\tilde{y}} + \rho_{\tilde{x}\tilde{y}} \sigma_{\tilde{y}} \frac{(\tilde{x} - \mu_{\tilde{x}})}{\sigma_{\tilde{x}}}, \quad (7)$$

and standard deviation

$$\sigma_{\tilde{y}|\tilde{x}, \theta} = \sigma_{\tilde{y}}^2 (1 - \rho_{\tilde{x}\tilde{y}}^2), \quad (8)$$

where $\rho_{\tilde{x}\tilde{y}}$ is a correlation coefficient.

We draw $N = 1000$ random samples from (6) to produce an ensemble forecast \tilde{x}^* . This forecast is then back-transformed to produce the calibrated ensemble forecast x^* . (Note that we use $*$ more generally to denote calibrated ensemble forecasts in the equations that follow.)

4) REORDERING CALIBRATED FORECASTS

The calibration produces an ensemble forecast at each location and lead time but does not link these forecasts in space or time. We instill spatial and temporal properties in each forecast using the Schaake shuffle ([Clark et al. 2004](#)).

The procedure is as follows. We begin with an empty matrix \mathbf{z} with elements $z_{i,\tau,t}$, where i is an index of locations, τ_h is an index of forecast lead time with dimension length $L = 36$ h, and t is an index of observed dates. The t dimension has length $T = N = 1000$, corresponding to the number of ensemble members. We then fill this matrix with rainfall sequences from the historical record. For example, we put a sequence of observations at a given location i and starting at time t in the matrix: $\mathbf{z}_{i,\bullet,t} = (z_{i,t+1,t} \ z_{i,t+2,t} \ \dots \ z_{i,t+L,t})$ where the notation \bullet denotes the i th row. We then add the sequence for this same time period at each location and repeat the whole process for T different starting dates. If the available T date sequences are less than the number of members N , then we repeat the shuffling process K times, sampling L sequences of observations from the T available sequences, such that $L < T$ and $L \times K = N$.

Once we have filled \mathbf{z} , we sort along the t dimension for each location i and lead time τ :

$$\begin{aligned}\hat{\mathbf{z}}_{i,\tau,\bullet} &= [z_{i,\tau,(1)} \ z_{i,\tau,(2)} \ \dots \ z_{i,\tau,(T)}] \\ z_{i,\tau,(1)} &\leq z_{i,\tau,(2)} \leq \dots \leq z_{i,\tau,(T)},\end{aligned}\quad (9)$$

with tied values (e.g., caused by the presence of zeros in the data) in \mathbf{z} assigned randomized ranks in $\hat{\mathbf{z}}$ in the t dimension. As with all applications of the Schaake Shuffle, the reordering process can be sensitive to the number of zeros present in the dependence template data ([Bellier et al. 2017](#)). We define an index matrix \mathbf{r} of elements $r_{i,\tau,t}$, at each location i and lead time τ by

$$\mathbf{z}_{[\mathbf{r}]} = \hat{\mathbf{z}}. \quad (10)$$

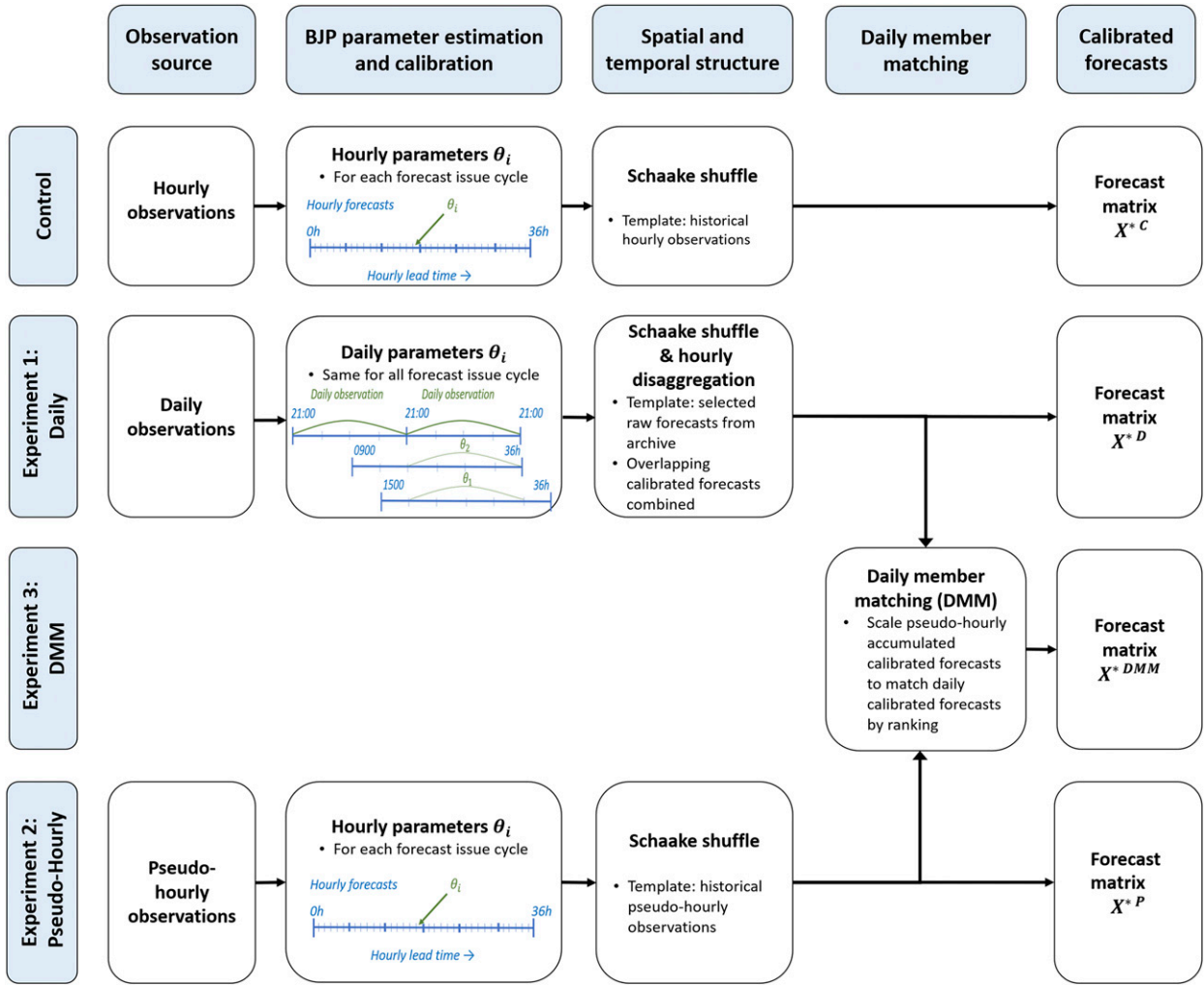


FIG. 2. Key methodological steps to generate calibrated ensemble forecasts with the control case and experiments 1, 2, and 3.

The index matrix \mathbf{r} describes rank correlations in space and time in \mathbf{z} . This is the template from which we reorder our forecast.

After calibration we have an ensemble forecast at each location i and each lead time τ :

$$\mathbf{x}_{i,\tau,\bullet}^* = (x_{i,\tau,1}^*, x_{i,\tau,2}^*, \dots, x_{i,\tau,N}^*), \quad (11)$$

where $N = 1000$ is the size of the ensemble. Note that \mathbf{x}^* and \mathbf{z} are of identical size. As with Eq. (9), we sort each \mathbf{x}^* along the n dimension for each location i and lead time τ :

$$\begin{aligned} \hat{\mathbf{x}}_{i,\tau,\bullet}^* &= [x_{i,\tau,(1)}^*, x_{i,\tau,(2)}^*, \dots, x_{i,\tau,(N)}^*] \\ x_{i,\tau,(1)}^* &\leq x_{i,\tau,(2)}^* \leq \dots \leq x_{i,\tau,(N)}^*. \end{aligned} \quad (12)$$

Tied values in \mathbf{x}^* (e.g., zeros) are assigned randomized ranks in $\hat{\mathbf{x}}^*$. The index matrix \mathbf{r} [Eq. (B2)] is then

used to reorder $\hat{\mathbf{x}}^*$ to produce a shuffled, calibrated forecast:

$$\mathbf{x}^{*SS} = \hat{\mathbf{x}}_{[\mathbf{r}]}^*. \quad (13)$$

b. Experimental design

To establish a new method to calibrate hourly precipitation forecasts from daily observations, we conduct three experiments as summarized in Fig. 2. We compare these to a control experiment, which represents the ideal case. The details are as follows.

1) CONTROL CALIBRATION

The control calibration represents the ideal case where high-quality hourly observations are available. Accordingly, forecasts are calibrated and shuffled with

the method described in [section 3a](#) (with details in [appendix A](#)) using hourly precipitation observations.

2) EXPERIMENT 1: DAILY CALIBRATION

Assuming that observed hourly data are unavailable, we cannot rely on hourly data to inform our calibration. This experiment thus differs from the control in three ways. First, forecasts are calibrated at a daily time step instead of the hourly time step. Second, the Schaake shuffle uses the raw NZCSM hourly deterministic forecasts as template data for reordering the ensemble forecast. Third, calibrated daily forecasts are disaggregated to hourly as part of the shuffling process. The process is as follows.

(i) Daily parameter estimation and calibrated forecasts

In this study, we have 36-h forecast length and daily observations aggregated over the hours 2100–2100 UTC. This creates an obvious problem: how do you calibrate a 36-h forecast using 24-h aggregations? A solution is to match the aggregation period of observations with separate forecast issue cycles. This allows us to calibrate two periods: the first 24 h of the forecast (lead 1–24) and the last 24 h (lead 13–36) of the forecast. We then must decide which calibration period to use on the overlapping interval (lead 13–24): we use the calibration of the first 24 h for this period. The procedure is as follows.

We infer calibration parameters from the 2100 cycle and 0900 cycle forecasts, which correspond to the first and last 24-h period of each forecast. For a given cycle and location, denote the archive of NZCSM hourly forecasts by the matrix,

$$\mathbf{x} = \begin{pmatrix} x_{1,1} & \cdots & x_{1,L} \\ \vdots & x_{t,\tau_h} & \vdots \\ x_{T,1} & \cdots & x_{T,L} \end{pmatrix}, \quad (14)$$

with dimensions $T \times L$ where $T = 1157$ forecasts and $L = 36$, as described in [section 2b](#).

To calibrate the 2100 cycle forecasts, we sum forecasts in \mathbf{x} along the lead time (τ_h) dimension to create the vector

$$\mathbf{D}^{2100} = \begin{pmatrix} \sum_{\tau_h=1}^{24} x_{1,\tau_h} \\ \vdots \\ \sum_{\tau_h=1}^{24} x_{T,\tau_h} \end{pmatrix}. \quad (15)$$

We then calibrate \mathbf{D}^{2100} forecasts to daily observations to generate the parameter set θ_{D1} , following the calibration method in [section 3a](#). To calibrate 0900 cycle

forecasts, we sum forecasts in \mathbf{x} along the lead time (τ_h) dimension to create the vector

$$\mathbf{D}^{0900} = \begin{pmatrix} \sum_{\tau_h=13}^{36} x_{1,\tau_h} \\ \vdots \\ \sum_{\tau_h=13}^{36} x_{T,\tau_h} \end{pmatrix}. \quad (16)$$

As with \mathbf{D}^{2100} , we calibrate \mathbf{D}^{0900} forecasts to daily observations to generate the parameter set θ_{D2} , following [section 3a](#).

Even though the two parameter sets are generated from the 2100 and 0900 cycles, we apply the parameters to all cycles. The method presumes forecasts from different cycles will have similar properties at similar lead times (first 24 h and last 24 h). To do this, we must first sum our deterministic hourly forecasts into daily totals. For each cycle we produce the matrix \mathbf{D} , as follows:

$$\mathbf{D} = \begin{pmatrix} \sum_{\tau_h=1}^{24} x_{1,\tau_h} & \sum_{\tau_h=13}^{36} x_{1,\tau_h} \\ \vdots & \vdots \\ \sum_{\tau_h=1}^{24} x_{T,\tau_h} & \sum_{\tau_h=13}^{36} x_{T,\tau_h} \end{pmatrix}. \quad (17)$$

Matrix \mathbf{D} has dimensions $T \times L_D$ where $L_D = 2$ lead time. The calibrated forecasts are unusual in that the summation periods for the two lead times overlap (for lead times of 13–24 h in \mathbf{x}). As noted above, this is necessary to enable the full 36-h of forecasts in \mathbf{x} to be calibrated. We generate calibrated ensembles from \mathbf{D} by applying θ_{D1} to $\tau_D = 1$ and θ_{D2} to $\tau_D = 2$. For each issue cycle t , this results in a matrix of calibrated ensemble forecasts \mathbf{D}^* of elements D_{n,τ_D}^* , where the n dimension is the ensemble size and is of length $N = 1000$. Note that \mathbf{D}^* is also used in the third experiment, described later in [section 3b\(4\)](#).

(ii) Ensemble reordering and hourly disaggregation

To produce spatially and temporally structured hourly ensemble members, we use the Schaake shuffle. However, as we assume hourly observations are not available for this experiment, we construct our template data from the hourly raw forecasts \mathbf{x} . Both ranking and forecast hourly disaggregation are assembled at the same time to produce a matrix \mathbf{x}^{*D} of calibrated forecasts with dimensions of lead time (36 h) and ensemble size $N = 1000$. The procedure is detailed in [appendix B](#).

We repeat the entire procedure of Eqs. (B1)–(B9) in appendix B for each forecast in **D**, to produce a set of calibrated, shuffled, and disaggregated forecasts at each gauge and for each cycle.

Note that we carry out this procedure in such a way that spatial and temporal patterns from the NZCSM forecasts are implicitly retained in the calibrated forecasts.

3) EXPERIMENT 2: PSEUDOHOURLY CALIBRATION

As with the experiment 1, for experiment 2 we assume only daily observations are available for calibration. But instead of calibrating forecasts to daily observations and then applying a disaggregation, we first synthesize hourly “observations” (termed *pseudohourly* observations) and then calibrate the forecasts to the pseudohourly observations. To generate pseudohourly observations, we disaggregate daily observations with temporal and spatial patterns from the NWP (Fig. 3). The disaggregation follows these steps:

- 1) Each daily observation, $\mathbf{y}_{i,t}$ at location i and time t , is matched to a corresponding raw forecast $\mathbf{x}_{i,t}$ that covers the same 24-h period:

$$\mathbf{x}_{i,t} = (x_{i,t-24} \ x_{i,t-23} \cdots x_{i,t-1}). \quad (18)$$

- 2) A set of weights is calculated for all forecast lead times. Weights are based on forecast patterns if daily rainfall totals are strictly positive, or based on even weights otherwise:

$$\mathbf{w}_{i,t} = \begin{cases} \left(\frac{x_{i,t-24} \ x_{i,t-23} \cdots x_{i,t-1}}{\sum_{\tau_h=1}^{24} x_{i,t-\tau_h}} \right), & \sum_{\tau=1}^{24} x_{i,t-\tau_h} > 0 \\ \frac{\mathbf{1}_{24}}{24}, & \sum_{\tau=1}^{24} x_{i,t-\tau_h} = 0 \end{cases}, \quad (19)$$

where $\mathbf{1}_{24}$ is the vector of all ones with 24 elements.

- 3) Twenty-four pseudohourly observations are calculated by multiplying the daily observation by the matrix of weights:

$$\mathbf{y}_{i,t}^p = \mathbf{y}_{i,t} \mathbf{w}_{i,t}. \quad (20)$$

We then use our pseudohourly observations \mathbf{y}^p to calibrate and shuffle our forecasts, as described in section 3a.

Matching a forecast to each daily observation (step 1) is complicated by the availability of forecasts from multiple

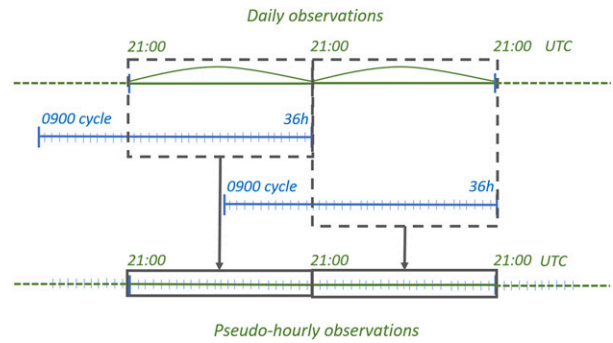


FIG. 3. Hourly disaggregation process of daily observations to generate pseudohourly observations; hourly temporal patterns from the raw forecasts 0900 cycle are used here.

cycles. In choosing which forecast cycle to use for the disaggregation, we do not wish to produce pseudohourly observations that match the timing of rainfalls in the raw forecasts too closely. If correlations between forecasts and pseudohourly observations are unrealistically high, this will cause the BJP to underestimate the true uncertainty in the forecast. For example, if we calibrate forecasts issued for the 1500 cycle against pseudo observations disaggregated from forecasts from the 1500 cycle (i.e., the same forecasts) this will result in much higher correlations between forecasts and the pseudo observations (and hence our calibrated ensemble would be too narrow) than would be expected if we calibrated forecasts against gauged observations. Thus, we must calibrate forecasts from a given cycle against pseudo observations disaggregated with forecasts from a different cycle. Given these constraints, we first choose NWP forecasts issued at the 0900 cycle to produce pseudohourly observations. We use these pseudohourly observations to calibrate forecast cycles 0300, 1500, and 2100. To calibrate the 0900 cycle forecasts, however, we use pseudohourly observations disaggregated with the 1500 cycle.

4) EXPERIMENT 3: DAILY MEMBER MATCHING CALIBRATION (DMM)

As we will show, experiment 2 generally produced more reliable hourly forecasts than experiment 1. Conversely, experiment 1 produced more reliable forecasts of 24-h rainfall totals. In this experiment, we wish to combine the best aspects of experiments 1 and 2. To do this, daily accumulated rainfall forecasts from the pseudohourly and daily methods are ranked, matched, and then scaled. The procedure is as follows.

For a given forecast cycle, location, and issue cycle, \mathbf{x}^{*p} denotes a forecast calibrated with the pseudohourly

method, of dimensions $N \times L$. We construct accumulated forecasts for \mathbf{x}^{*P} :

$$\mathbf{D}^{*P} = \begin{pmatrix} \sum_{\tau_h=1}^{24} x_{1,\tau_h}^{*P} & \sum_{\tau_h=13}^{36} x_{1,\tau_h}^{*P} \\ \vdots & \vdots \\ \sum_{\tau_h=1}^{24} x_{N,\tau_h}^{*P} & \sum_{\tau_h=13}^{36} x_{N,\tau_h}^{*P} \end{pmatrix}, \quad (21)$$

where \mathbf{D}^{*P} is of dimensions $N \times L_D$. We also retrieve \mathbf{D}^* from the daily method [see [section 3b\(2\)](#)], for the same cycle, location, and issue time. The matrix \mathbf{D}^* contains forecasts of 24-h accumulations that have been calibrated to daily observations.

Next, we sort \mathbf{D}^{*P} and \mathbf{D}^* along the n dimension, to produce sorted matrices $\hat{\mathbf{D}}^{*P}$ and $\hat{\mathbf{D}}^*$, respectively, following Eq. (9). We define an index matrix \mathbf{r} to enable us to reverse the sorting process in the pseudohourly forecasts by $\hat{\mathbf{D}}_{[\mathbf{r}]}^{*P} = \mathbf{D}^{*P}$.

We then compute scale factors to scale the pseudohourly forecast accumulations to match the daily accumulations in \mathbf{D}^* . For each ensemble member, $n = 1, 2, \dots, 1000$ and each lead time $\tau_D = 1, 2$, the scale factor is calculated by

$$\hat{\mathbf{A}}_{n,\tau_D} = \begin{cases} \frac{\hat{\mathbf{D}}_{n,\tau_D}^*}{\hat{\mathbf{D}}_{n,\tau_D}^{*P}}, & \hat{\mathbf{D}}_{n,\tau_D}^{*P} \geq \varepsilon \\ 1, & \hat{\mathbf{D}}_{n,\tau_D}^{*P} < \varepsilon \end{cases}, \quad (22)$$

where ε is a small positive number to avoid division by near-zero rainfalls. We tested different thresholds for $\varepsilon = 0.05, 0.1$, and 0.5 mm. We found that $\varepsilon = 0.05$ mm was the smallest threshold that maximized the number of forecasts to be scaled by the daily forecasts, while avoiding divisions by 0. We unsort the scaling factors with the index matrix \mathbf{r} by $\mathbf{A} = \hat{\mathbf{A}}_{[\mathbf{r}]}$. Scaled pseudohourly forecasts are then calculated for each ensemble member by

$$\mathbf{x}_{n,\bullet}^{*DMM} = \left[\mathbf{A}_{n,1} \left(x_{n,1}^{*P} x_{n,2}^{*P} \cdots x_{n,24}^{*P} \right) \quad \mathbf{A}_{n,2} \left(x_{n,25}^{*P} x_{n,26}^{*P} \cdots x_{n,36}^{*P} \right) \right]. \quad (23)$$

Hourly forecasts $\mathbf{x}_{n,\bullet}^{*DMM}$ denote the τ_h th row of \mathbf{x}^{*DMM} . The result is pseudohourly calibrated forecasts that have rainfall accumulations consistent with forecasts from the daily method. This process [Eqs. (21)–(23)] is repeated for each cycle and location.

c. Forecast verification by cross validation

We assess three key performance aspects of the ensemble rainfall forecasts: errors, bias, and reliability. Errors show the accuracy of the forecast, while bias indicates a general tendency to over or underpredict observations. Reliability indicates the appropriateness of the ensemble spread—i.e., ensemble spread is correctly distributed, and not too wide or too narrow. Bias is often considered a component of reliability—usually biased forecasts are not reliable. However, in the case of a highly skewed variables such as rainfall, a few outlying values can cause strong bias while forecasts can be reliable overall.

For each method and each station, the performance of rainfall forecasts is assessed against observed station data (available hourly). Performance is assessed at individual lead times, and for cumulative totals with 12-h accumulations (lead time 1–12, 13–24, and 25–36), with 24-h accumulations (lead time 1–24), and with 36-h accumulations (lead time

1–36). This enables us to independently assess the univariate calibration method and the reordering ensemble generation method.

We use a leave-one-month-out cross-validation procedure to ensure that the forecasts are verified independently of model fitting. For all methods, Bayesian joint probability parameters are inferred using all available data except one month. All the forecasts in that left-out month are then verified to the corresponding hourly station observations.

1) FORECAST RELIABILITY

We check forecast reliability with the probability integral transform (PIT). Given the cumulative distribution function of a forecast F_t , the PIT is given by

$$\text{PIT} = \begin{cases} F_t(y), & y(t) > 0 \\ U(0, 1) \times F_t(0), & y(t) = 0 \end{cases}. \quad (24)$$

For a reliable set of forecasts, PIT values should be uniformly distributed. The treatment of PIT values at $y(t) = 0$ is necessary to allow reliable predictions to produce uniformly distributed PIT values when zero rainfalls occur ([Wang and Robertson 2011](#)).

We check uniformity by plotting PIT values as histograms. We calculate PIT values for individual lead times

Hourly PIT histogram, all sites

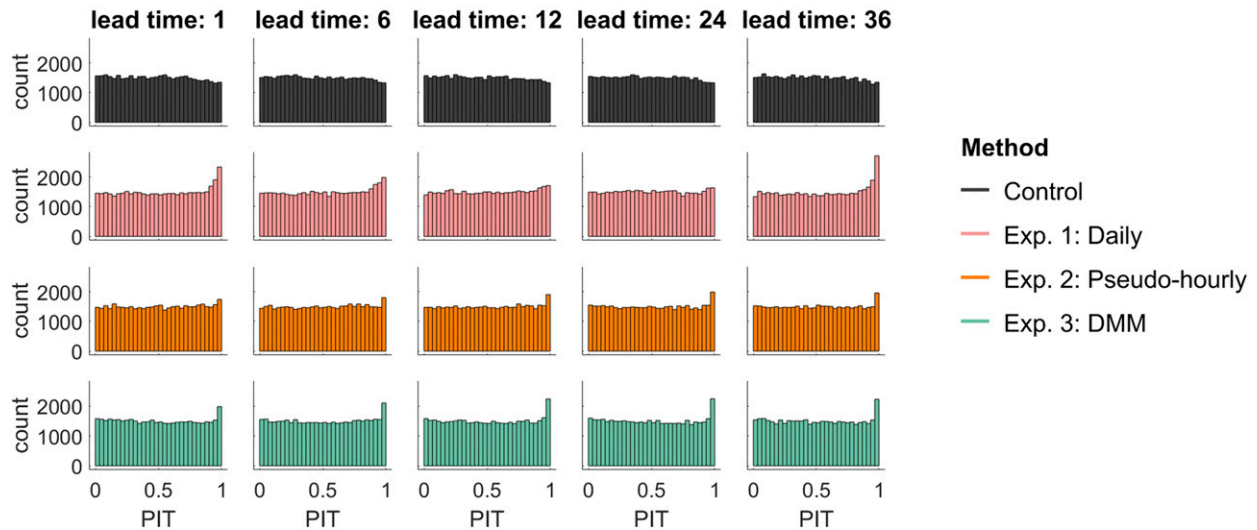


FIG. 4. Hourly PIT histograms for each method and all sites as a function of selected lead times.

and for accumulated rainfalls. Forecasts of accumulated rainfalls can only be reliable if the ensemble has realistic spatial and temporal patterns.

2) FORECAST BIAS

We measure forecast bias with relative bias:

$$\text{bias} = \frac{\bar{x} - \bar{y}}{\bar{y}} \times 100\%, \quad (25)$$

where \bar{x} (or \bar{x}^* , in the case of ensemble forecasts) is the mean of a set of forecasts and \bar{y} is the mean of the corresponding set of observations.

3) FORECAST ACCURACY

We measure errors in probabilistic forecasts with the continuous ranked probability score (CRPS). For a set of forecasts at $t = 1, 2, \dots, T$,

$$\text{CRPS} = \frac{1}{T} \sum_{t=1}^T \int_{-\infty}^{\infty} \{F_t(z) - H[y(t) \leq z]\}^2 dz, \quad (26)$$

where F_t is the cumulative distribution function (CDF) of the forecast distribution, and H is the Heaviside step function.

CRPS reduces to the mean absolute error (MAE) for deterministic predictions, allowing us to compare errors in uncalibrated deterministic forecasts to errors in calibrated probabilistic forecasts. CRPS is negatively oriented: smaller scores indicate better forecasts, with zero being a perfect forecast. We use bootstrap resampling to assess the significance of reduction in CRPS error relative to the raw forecasts.

4. Results

a. Forecast reliability

Figure 4 presents PIT histograms calculated for all gauge stations at individual lead times. The control method generates PIT histograms that are close to ideal. The pseudohourly forecasts produce a slight peak to the right of the histogram, indicating a faint negative bias. This negative bias is exacerbated somewhat in the daily forecasts. The DMM method produces results that combine aspects of the pseudohourly and daily methods: largely reliable, with slight evidence of negative bias. Slight overpopulation in the last bin of the hourly histograms (Fig. 4) can be observed for all sites; these are due to missed rainfall events from inaccurate timing or underestimation in forecast rainfall. We note that poor NWP timing can be amplified in our method because both the pseudohourly observations and forecasts rely on the timing of NWP forecast rainfall being right.

Figure 5 presents PIT histograms of rainfall accumulations for selected 12-, 24-, and 36-h totals. The daily method produces forecast accumulations that are almost perfectly reliable, especially for the 24-h periods. Conversely, PIT histograms for the control and pseudohourly method deviate from the horizontal line with lower counts for higher PIT values. Scaling pseudohourly calibrated forecasts using daily calibrated forecasts significantly improves reliability of accumulations; PIT histograms from the DMM method display nearly flat histograms compared to those of the pseudohourly method.

We believe the imperfect reliability of accumulated forecasts from the control and pseudohourly methods to

Accumulated PIT histogram, all sites

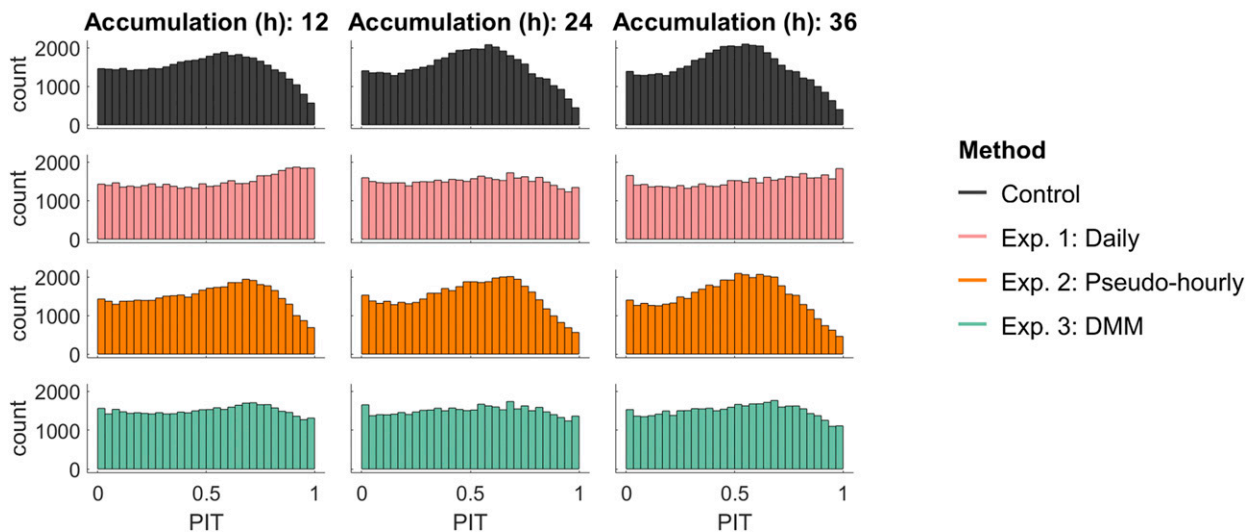


FIG. 5. Accumulated PIT histograms for each method and all sites for 12-, 24-, and 36-h totals.

be related to the mismatch in autocorrelation of gauged rainfall and raw NWP forecasts. NWP forecasts tend to vary much more smoothly in time than gauged rainfalls, i.e., NWP forecasts are more autocorrelated than gauged observations. This appears to cause rainfall accumulations to be underconfident even though forecasts at individual lead times are highly reliable. We have identified this issue in other work and are currently investigating the exact cause. In previous applications (Bennett et al. 2014; Robertson et al. 2013; Shrestha et al. 2015), the BJP calibration was applied to areally averaged rainfalls, which tend to exhibit similar autocorrelation to NWP rainfall. In these previous studies, calibrated forecasts of accumulated rainfalls were highly reliable. As the aim of this present study is to generate a method that will ultimately be applied to an areally averaged rainfall product (VCSN) rather than to gauged rainfalls, this is not a major failing of the method applied here. We also note that the DMM largely resolves this issue by matching ensemble daily precipitation totals to values calibrated to daily observations.

Additionally, the spatial dependence structure of ensemble forecasts is preserved in the DMM method with the Schaake shuffle [section 3a(4) and appendix B]. This is illustrated by the PIT plots of spatial catchment average for both hourly lead times and accumulations of 12, 24, and 36 h, provided in the online supplemental material (Figs. S1 and S2).

b. Forecast bias

We assess forecast bias for each method by presenting boxplots of the mean forecast bias values over the different sites.

Figure 6 presents hourly bias in the raw NWP and calibrated forecasts. Calibrated forecasts have markedly smaller bias than the raw forecasts. The control method displays little bias (close to zero) at all lead times. This is to be expected: by construction the BJP method optimizes parameters to produce unbiased forecasts. Pseudohourly forecasts tend to be positively biased, although the biases are reasonably small, particularly in contrast to the daily method. The daily method produces forecasts that exhibit strong bias for all sites, with up to 40% negative bias at early and late lead times and up to 40% positive bias around lead time 18–20 h. This is because the calibration minimizes bias at the daily time step but is given no information to minimize biases at subdaily lead times. Note that the strong bias sometimes evident in the daily method (Fig. 6) does not always manifest strongly in the PIT histograms (Fig. 4). This is because of the strongly skewed nature of rainfall: a small number of very large differences between observations and forecasts are sufficient to cause large biases (Fig. 6). However, because these instances are few, they are not strongly evident in the PIT histograms (Fig. 4). The DMM method produces smaller biases than either the pseudohourly or daily method, and biases are fairly consistent across all lead times.

Figure 7 presents bias of rainfall accumulations for the raw and calibrated forecasts. Overall bias of accumulated forecasts is smaller than bias for individual lead times. This is because errors at individual lead times tend to compensate for each other in the accumulated totals. Calibrated forecast bias is significantly

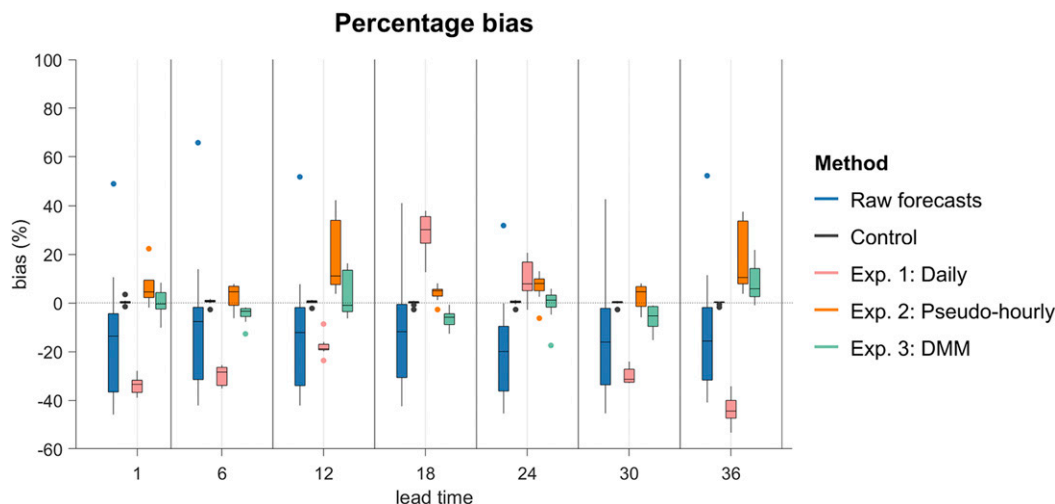


FIG. 6. Hourly relative bias for the deterministic NWP and postprocessed forecasts, for each method and across all sites at lead times 1, 6, 12, 18, 24, 30, and 36 h. Unbiased forecasts lie along the dashed line. The box is drawn between the 25th and 75th percentiles, with a line indicating the median. The whiskers extend above and below the box to the most extreme data points that are within a distance to the box equal to 1.5 times the interquartile range (Tukey boxplot). Points outside the whisker ranges are plotted.

smaller than raw forecast bias for 12-, 24-, and 36-h accumulations. The smallest accumulated biases across all sites are for forecasts from the control method. These are centered around zero and have a narrow spread across sites. Forecasts using the pseudohourly method consistently overforecast accumulated precipitation by 10%. The daily calibration produces forecasts with bias centered around zero for the 24-h accumulation. This is expected, as this method calibrates forecasts directly to daily data. The DMM method fulfils its objective by improving

the performance of the pseudohourly method: mean forecast bias of precipitation accumulation is small and centered around zero.

c. Forecast accuracy

We assess forecast accuracy for each method by presenting boxplots of the mean forecast CRPS and MAE values over the different sites.

Calibrated forecasts have substantially lower average errors than the raw NWP predictions at all sites

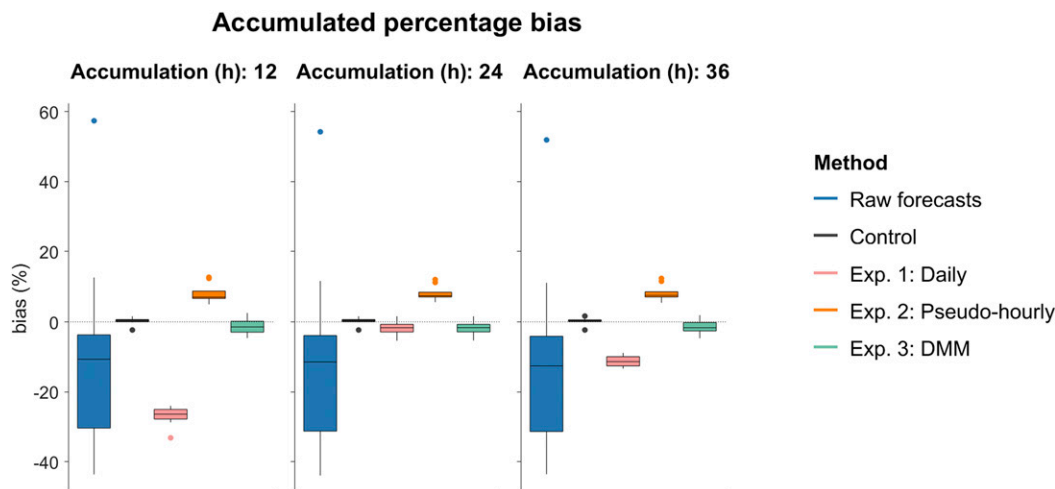


FIG. 7. Accumulated relative bias for the deterministic NWP and postprocessed forecasts, for each method and across all sites, as a function of 12-, 24-, and 36-h accumulations. Quantiles in boxplots are as for Fig. 6.

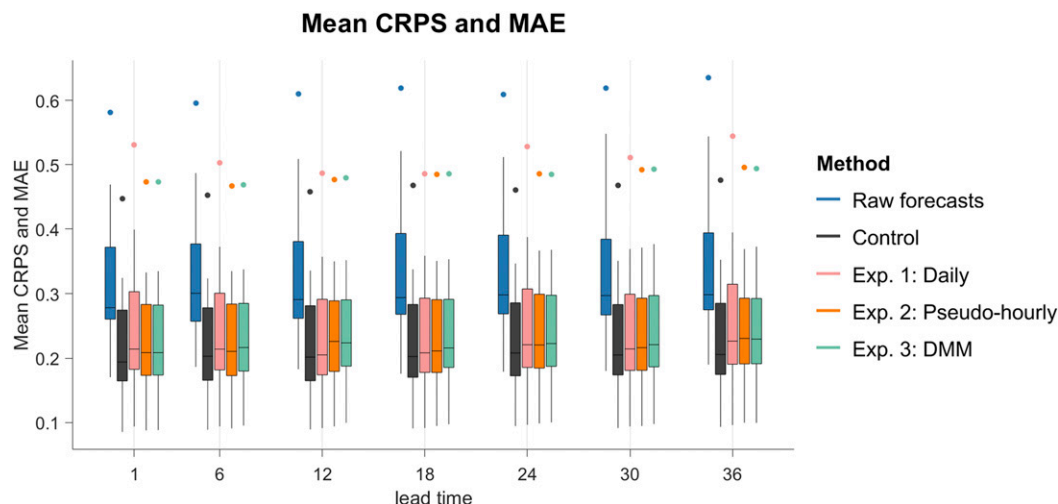


FIG. 8. Hourly MAE for raw deterministic NWP forecasts and CRPS for postprocessed forecasts for each method and across all sites, as a function of lead time 1, 6, 12, 18, 24, 30, and 36 h. Quantiles in boxplots are as for Fig. 6.

and lead times (Fig. 8). As expected, calibrated forecasts using the control method have the lowest errors, followed very closely by forecasts using the pseudo-hourly and DMM methods. The daily method produces the worst accuracy of the calibrated forecasts at individual lead times.

Figure 9 presents CRPS and MAE of accumulated rainfalls summarized for all sites. All calibration methods outperform raw forecasts for all sites, and all offer similar performance. Interestingly, forecasts based on the DMM method have the highest accuracy for the 36-h accumulation, though the difference in accuracy between all calibration methods is very small.

For both hourly and accumulated forecasts, all the calibration methods lead to statistically significant reductions in error relative to the raw forecasts. This is illustrated in Fig. S3 for the DMM method.

5. Discussion

The effectiveness of the daily member matching DMM method (experiment 3) is due to the combination of the best aspects of the daily method (experiment 1) and the pseudohourly method (experiment 2). The DMM method produces reliable and bias free accumulated forecasts (a property of the daily calibration) without a bias pattern at hourly lead times (a property of the pseudohourly calibration).

As with other postprocessing methods, the DMM calibration requires a reasonable size of template data and forecast archive, often a challenge due to limited

availability of a homogeneous NWP forecast archive. From a hydrological perspective, a 3-yr archive is a short record to establish a climatology of observed rainfall and space–time patterns for the Schaake shuffle. In addition, extreme rainfall may be missed, affecting the calibration for extreme events. Long reforecast archives are very valuable for detecting and correcting systematic errors in forecasts, especially forecasts of relatively rare events (Hamill et al. 2013). Longer reforecast archives also make it simple to generate longer records of template data for ensemble reordering, which better reflect the full historical range of spatiotemporal precipitation patterns.

A key assumption in the calibration method is that the NWP characterizes the spatial and temporal patterns of rainfall well. NWP spatial and temporal patterns underpin the pseudohourly observations, used as the “truth” to which forecasts are calibrated, as well as the basis of the Schaake shuffle. NWP models often differ from observations in crucial ways: for example, there may be a mismatch in diurnal patterns (Shrestha et al. 2015; Surcel et al. 2010). In these cases, the BJP method may overestimate true correlations between observations and forecasts, because the pseudohourly observations are much more like the forecasts than actual observations. This can lead the calibration to amplify overestimation or underestimation in the raw forecasts, causing the pseudohourly method to produce biases at individual lead times. Given these difficulties, we do not recommend the use of the pseudohourly method on its own. We reiterate, however, that the DMM method successfully mitigates these problems.

Accumulated mean CRPS and MAE

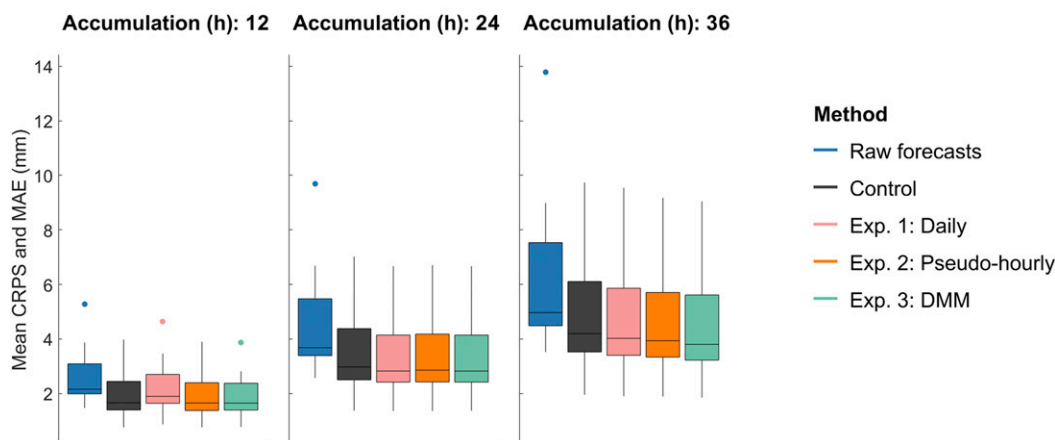


FIG. 9. Accumulated MAE for raw deterministic NWP forecasts and CRPS for postprocessed forecasts for each method and across all sites, as a function of 12-, 24-, and 36-h accumulations. Quantiles in boxplots are as for Fig. 6.

Future work could include using reordering methods with preferential selection of past observations having similar atmospheric states than current forecasts (Scheffzik 2016; Scheuerer et al. 2017). This could improve the predicted hourly temporal structure as the Schaake shuffle would be informed by a more representative sample of historic events. These could be stratified using meteorological analogs (Bellier et al. 2017) and citations therein) and could be particularly valuable when separating stratiform from convective precipitation as these have distinct temporal patterns. Although our calibration method is applied to a raw deterministic NWP, it could be applied to a raw ensemble NWP with the application of the ECC to preserve spatial and temporal dependency structure in calibrated forecasts (Scheffzik et al. 2013).

Joint or individual calibration of other variables (e.g., temperature) may be required for developing a national-scale flow forecasting system in New Zealand (Monhart et al. 2019). For example, snow and glacier melt is an important contributor to runoff in many rivers. A national-scale calibration approach may need special handling of distant station or grid points for the ensemble reordering aspect and may face computational constraints associated with a larger domain.

6. Summary

This study establishes a new method (daily member matching or DMM) to calibrate hourly precipitation forecasts from daily observations. The DMM method combines a daily calibration approach with an hourly

calibration approach using hourly forecast patterns, by matching daily ensemble forecast values. The method is evaluated for ten stations in a catchment in New Zealand with steep rainfall gradients. The method performs similarly well to an ideal case where hourly data are available: calibrated forecasts have much lower bias and substantially smaller errors than the raw forecasts. In addition, the method produces reliable forecasts at individual lead times and for forecasts of precipitation accumulations.

Generating a statistically calibrated ensemble forecast from deterministic NWP predictions and daily data is likely to be of significant benefit for the expansion of streamflow forecasting services. Deterministic forecasts are routinely available (in New Zealand and elsewhere) at subdaily time steps (sometimes even subhourly) while daily precipitation observation datasets are much more common than subdaily datasets, and often available over large domains (national or continental scales). Scarcity of hourly observations is a problem in many regions, and particularly in developing countries.

Reliable, accurate and bias-free forecasts of catchment-scale precipitation are required to produce useable streamflow forecasts. This study is an important step toward the development of a national scale flow forecasting system in New Zealand to support a range of emergency services and water managers.

Acknowledgments. The authors gratefully acknowledge various parties for assistance in providing data (in particular the Greater Wellington Regional Council). This research was funded by the New Zealand Ministry of Business, Innovation and Employment Natural Hazards

Research Platform under contract C05X0907/Subcontract 2017-NIW-03-NHRP; by NIWA through the Resilience to Hazards Research Programme. The authors wish to acknowledge the contribution of NeSI to the results of this research. New Zealand's national compute and analytics services and team are supported by the New Zealand eScience Infrastructure (NeSI) and funded jointly by NeSI's collaborator institutions and through the Ministry of Business, Innovation and Employment (<http://www.nesi.org.nz>). Figures in this paper were produced using the Gramm MATLAB package (Morel 2018). The editor and the two reviewers are gratefully acknowledged for their valuable and thoughtful feedback.

APPENDIX A

Parameter Estimation Procedure

Parameter estimation is carried out in stages. In the first stage, the transformation and marginal distribution parameters a, b, μ, σ are estimated separately for observations and forecasts. We will describe the procedure to estimate these parameters for forecasts. An identical estimation is carried out for observations. To ease inference, we reparameterize and infer $\theta_x = \{\log(a_x), \log(b_x), \mu_{\tilde{x}}/\sigma_{\tilde{x}}, \log(\sigma_{\tilde{x}})\}$. For a vector of forecasts scaled with Eq. (1), $\mathbf{x}' = \{x'_1, x'_2, \dots, x'_n\}$, the posterior distribution of θ_x is given by

$$p(\theta_x|\mathbf{x}') \propto p(\theta_x)p(\mathbf{x}'|\theta_x) = p(\theta_x)\prod_{t=1}^n p(\tilde{x}_t|\theta_x), \quad (\text{A1})$$

where $p(\theta_x)$ is the prior distribution and $p(\mathbf{x}|\theta_x)$ is the likelihood. The likelihood is given by

$$p(\mathbf{x}'|\theta_x) = \begin{cases} J_{\tilde{x} \rightarrow x'} N(\tilde{x}|\mu_{\tilde{x}}, \sigma_{\tilde{x}}^2), & \tilde{x} > \tilde{x}_c \\ \Phi(\tilde{x}_c|\mu_{\tilde{x}}, \sigma_{\tilde{x}}^2), & \tilde{x} \leq \tilde{x}_c \end{cases}, \quad (\text{A2})$$

where \tilde{x}_t is the log-sinh transform of x'_t [Eq. (2)], the Jacobian is

$$J_{\tilde{x} \rightarrow x'} = \frac{1}{\tanh[a_x + b_x \tilde{x}]}, \quad (\text{A3})$$

$\tilde{x}_c = b_x^{-1} \log[\sinh(a_x + b_x x'_c)]$ is the log-sinh transformed value of the scaled censoring threshold [Eq. (A2)] and $\Phi(\tilde{x}_c|\mu_{\tilde{x}}, \sigma_{\tilde{x}}^2)$ is the cumulative distribution function of a univariate normal distribution.

The prior is given by

$$p(\theta_x) = p(a_x)p(b_x)p(\mu_{\tilde{x}})p(\sigma_{\tilde{x}}), \quad (\text{A4})$$

where

$$\begin{aligned} p(a_x) &\propto 1 \quad a_x \leq 1, \\ p[\log(b_x)] &\propto N(0, 1^2), \\ p(\mu_{\tilde{x}}) &\propto 1, \\ p(\sigma_{\tilde{x}}) &\propto 1, \end{aligned} \quad (\text{A5})$$

The priors for a_x , $\mu_{\tilde{x}}$, and $\sigma_{\tilde{x}}$ are uninformative. We impose the restriction of $a_x \leq 1$ because for values of $a_x > 1$ the log-sinh transformation has little effect on the skewness of data. The prior on $\log(b_x)$ is informative, and encourages b_x to be close to 1 if the data are not strongly skewed.

The maximum posterior density of Eq. (A1) is found with the shuffled complex evolution (SCE) algorithm (Duan et al. 1992).

As noted above, the process of finding θ_x [Eqs. (A1)–(A5)] is repeated for observations. Once transformation and marginal distribution parameters are estimated for both observations and forecasts, $\theta_{x,y} = [a_{x'} \ b_{x'} \ \mu_{\tilde{x}} \ \sigma_{\tilde{x}} \ a_{y'} \ b_{y'} \ \mu_{\tilde{y}} \ \sigma_{\tilde{y}}]$ [i.e., after reversing the reparameterizations in Eq. (A1)], parameters in $\theta_{x,y}$ are fixed.

The second stage is to estimate the correlation parameter $\rho_{\tilde{x}\tilde{y}}$. The likelihood is more complex for the bivariate normal distribution [Eq. (3)], as described by (Robertson et al. 2013). To ease inference, we reparameterize $\rho_{\tilde{x}\tilde{y}}$ to

$$\varphi = \tanh^{-1}(\rho_{\tilde{x}\tilde{y}}) \quad (\text{A6})$$

to give $\theta = [\theta_{x,y} \ \varphi]$, where $\theta_{x,y}$ is fixed. The posterior density of θ is given by

$$p(\theta|\mathbf{y}', \mathbf{x}') \propto p(\theta)p(\mathbf{y}', \mathbf{x}'|\theta) = p(\theta)\prod_{t=1}^n p(\tilde{y}_t, \tilde{x}_t|\theta), \quad (\text{A7})$$

where $\mathbf{x}' = \{x'_1, x'_2, \dots, x'_N\}$ is a vector of scaled forecasts [Eq. (1)] that correspond to observations in \mathbf{y}' . Because of the presence of zeros in both observations and forecasts, the likelihood in Eq. (A7) must consider four cases:

$$p(\tilde{y}, \tilde{x}|\theta) = \begin{cases} J_{\tilde{y} \rightarrow y'} J_{\tilde{x} \rightarrow x'} p(\tilde{y}, \tilde{x}|\theta), & \tilde{y} > \tilde{y}_c, \tilde{x} > \tilde{x}_c \\ J_{\tilde{x} \rightarrow x'} p(\tilde{y} \leq \tilde{y}_c, \tilde{x}|\theta), & \tilde{y} \leq \tilde{y}_c, \tilde{x} > \tilde{x}_c \\ J_{\tilde{y} \rightarrow y'} p(\tilde{y}, \tilde{x} \leq \tilde{x}_c|\theta), & \tilde{y} > \tilde{y}_c, \tilde{x} \leq \tilde{x}_c \\ p(\tilde{y} \leq \tilde{y}_c, \tilde{x} \leq \tilde{x}_c|\theta), & \tilde{y} \leq \tilde{y}_c, \tilde{x} \leq \tilde{x}_c \end{cases}, \quad (\text{A8})$$

where the Jacobian for forecast values is given by

$$J_{\tilde{x} \rightarrow x'} = \frac{1}{\tanh[a_x + b_x \tilde{x}]}, \quad (\text{A9})$$

$\tilde{x}_c = b_x^{-1} \log[\sinh(a_x + b_x x'_c)]$ is the log-sinh transformed value of the scaled censor threshold x'_c for forecasts,

$$\begin{aligned}
p(\tilde{y} \leq \tilde{y}_c, \tilde{x} | \theta) &= p(\tilde{x} | \theta) \int_{-\infty}^{\tilde{y}_c} p(\tilde{y} | \tilde{x}, \theta) d\tilde{y}, \\
p(\tilde{y}, \tilde{x} \leq \tilde{x}_c | \theta) &= p(\tilde{y} | \theta) \int_{-\infty}^{\tilde{x}_c} p(\tilde{x} | \tilde{y}, \theta) d\tilde{x}, \\
p(\tilde{y} \leq \tilde{y}_c, \tilde{x} \leq \tilde{x}_c | \theta) &= \int_{-\infty}^{\tilde{y}_c} \int_{-\infty}^{\tilde{x}_c} p(\tilde{y}, \tilde{x} | \theta) d\tilde{y} d\tilde{x}, \quad (\text{A10})
\end{aligned}$$

and all other terms are as defined earlier. We do not impose an informative prior on $\rho_{\tilde{y}\tilde{x}}$ (i.e., $p(\rho_{\tilde{y}\tilde{x}}) \propto 1$), but the prior in Eq. (A7) must account for the reparameterization, as follows:

$$p(\theta) = p(\varphi) = J_{\varphi \rightarrow \rho_{\tilde{y}\tilde{x}}} p(\rho_{\tilde{y}\tilde{x}}) = J_{\varphi \rightarrow \rho_{\tilde{y}\tilde{x}}} [\cosh(\varphi)]^{-2}. \quad (\text{A11})$$

As with the transformation and marginal distribution parameters, we maximize the posterior density [Eq. (A7)] using the SCE algorithm.

APPENDIX B

Daily Calibration: Ensemble Reordering and Hourly Disaggregation

The daily calibration produces a matrix of daily accumulated and calibrated ensemble forecasts \mathbf{D}^* [section 3b(2)]. Here we describe the ensemble reordering and hourly disaggregation process, which are implemented concurrently. The conventional Schaake Shuffle uses observations as template data [section 3a(4)] but we assume hourly observations are not available for this experiment. To simplify index notation, we omit indices referring to station locations that were explicitly detailed in the Schaake shuffle in section 3a(4).

We therefore construct our template data from the hourly forecasts \mathbf{x} . Our first step in assembling the template data is to exclude some forecasts in \mathbf{x} . A given forecast at t is excluded if $\sum_{\tau_h=3}^{22} x_{t,\tau_h} < 0.4$ mm or if $\sum_{\tau_h=15}^{34} x_{1,\tau_h} < 0.4$ mm. These exclusions ensure that some rain occurs in the middle of each 24-h summation period in \mathbf{D} [defined in Eq. (17)]. This is necessary for the disaggregation, otherwise the beginning (and end) of each 24-h period is overrepresented in the template forecasts because rainfall there often corresponds to the very end (or beginning) of a rainfall event. Consequently, if rain falls only at the very beginning or the very end of either 24-h period, the disaggregation can assign unrealistically large rainfalls in only 1 or 2 h. After the exclusion, we are left with slightly more than 250 forecasts (e.g., 253; the exact value depends on each cycle), and we reduce this to exactly $T = 250$ by randomly removing excess forecasts.

We are now left with a subset of forecasts \mathbf{z} , which we sum to produce daily totals to generate the matrix \mathbf{Z} of dimensions $T \times L_D$:

$$\mathbf{Z} = \begin{pmatrix} Z_{1,1} & Z_{1,2} \\ \vdots & \vdots \\ Z_{T,1} & Z_{T,2} \end{pmatrix} = \begin{pmatrix} \sum_{\tau_h=1}^{24} z_{1,\tau_h} & \sum_{\tau_h=13}^{36} z_{1,\tau_h} \\ \vdots & \vdots \\ \sum_{\tau_h=1}^{24} z_{T,\tau_h} & \sum_{\tau_h=13}^{36} z_{T,\tau_h} \end{pmatrix}, \quad (\text{B1})$$

where \mathbf{Z} constitutes our template data for the Schaake shuffle and the hourly disaggregation. The uppercase notation \mathbf{Z} refers to accumulated forecast values and the lowercase notation \mathbf{z} refers to hourly forecast values. Following section 3a(4), for each lead time τ_D , we sort \mathbf{Z} along the T dimension to give

$$\hat{\mathbf{Z}}_{\bullet, \tau_D} = \begin{bmatrix} Z_{(1), \tau_D} \\ Z_{(2), \tau_D} \\ \vdots \\ Z_{(T), \tau_D} \end{bmatrix} \quad Z_{(1), \tau_D} \leq Z_{(2), \tau_D} \leq \dots \leq Z_{(T), \tau_D}. \quad (\text{B2})$$

Accumulated forecasts $\mathbf{Z}_{\bullet, \tau_D}$ denote the τ_D th column of \mathbf{Z} of dimensions $(T = 1157) \times (L_D = 2)$. We define two index matrices \mathbf{r}_{τ_D} , one for each lead time, by

$$\hat{\mathbf{Z}}_{[\mathbf{r}]} = \left(\hat{\mathbf{Z}}_{\bullet, 1[\mathbf{r}_1]} \quad \hat{\mathbf{Z}}_{\bullet, 2[\mathbf{r}_2]} \right) = \left(\mathbf{Z}_{\bullet, 1} \quad \mathbf{Z}_{\bullet, 2} \right) = \mathbf{Z}, \quad (\text{B3})$$

where \mathbf{r} is composed of the index vectors \mathbf{r}_1 and \mathbf{r}_2 , which map the unsorted forecasts in \mathbf{Z} to the sorted values in $\hat{\mathbf{Z}}$ for lead times $\tau_D = 1$ and $\tau_D = 2$, respectively. Following the Schaake shuffle [see Eq. (13) in section 3a(4)], \mathbf{r}_1 and \mathbf{r}_2 are used to reorder the forecast \mathbf{D}^* . As the ensemble size $N = 1000$ is larger than the available template data size, we sample a first set of 250 members from \mathbf{D}^* to reorder and disaggregate forecasts. For each τ_D , we sort \mathbf{D}^* along the row dimension with

$$\mathbf{D}_{\bullet, \tau_D}^* = \begin{bmatrix} D_{(1), \tau_D}^* \\ D_{(2), \tau_D}^* \\ \vdots \\ D_{(T), \tau_D}^* \end{bmatrix} \quad D_{(1), \tau_D}^* \leq D_{(2), \tau_D}^* \leq \dots \leq D_{(T), \tau_D}^*. \quad (\text{B4})$$

We disaggregate \mathbf{D}^* to the hourly time step using the index matrix \mathbf{r} to match the hourly forecasts accumulated in

\mathbf{Z} to the sorted 24-h accumulations in $\hat{\mathbf{Z}}$ [Eq. (10)]. That is, the ranked ensemble members in $\hat{\mathbf{D}}^*$ will be disaggregated to the hourly forecast patterns from \mathbf{z}

covering lead times 1–24 based on the $\hat{\mathbf{Z}}_{\bullet,1}$ ranking, while \mathbf{z} covering lead times 25–36 is based on patterns from the $\hat{\mathbf{Z}}_{\bullet,2}$ ranking:

$$\mathbf{z}_{[r]} = (\mathbf{z}_{[r_1],1 \leq \tau_h \leq 24} \quad \mathbf{z}_{[r_2],25 \leq \tau_h \leq 36}) = \begin{bmatrix} z_{(1)_1,1} & \cdots & z_{(1)_1,24} & z_{(1)_2,25} & \cdots & z_{(1)_2,36} \\ \vdots & & \vdots & \vdots & & \vdots \\ z_{(T)_1,1} & \cdots & z_{(T)_1,24} & z_{(T)_2,25} & \cdots & z_{(T)_2,36} \end{bmatrix}. \quad (\text{B5})$$

We now have hourly forecasts \mathbf{z} (lowercase) to a forecast horizon of 36-h lead time. For each forecast and hourly lead time, we can calculate a weight w_{t,τ_h} (hourly rainfall divided by the daily rainfall) given by

$$w_{t,1 \leq \tau_h \leq 24} = \frac{z_{[r_1],t,1 \leq \tau_h \leq 24}}{\sum_{\tau_h=1}^{24} z_{[r_1],t,\tau_h}}; \quad w_{t,25 \leq \tau_h \leq 36} = \frac{z_{[r_2],t,25 \leq \tau_h \leq 36}}{\sum_{\tau_h=25}^{36} z_{[r_2],t,\tau_h}}. \quad (\text{B6})$$

We then multiply the calibrated daily totals by the weights to produce ranked calibrated hourly forecasts:

$$\hat{\mathbf{X}}_{t,1 \leq \tau_h \leq 24}^* = w_{t,1 \leq \tau_h \leq 24} \hat{\mathbf{D}}_{t,1}^*; \quad \hat{\mathbf{X}}_{t,25 \leq \tau_h \leq 36}^* = w_{t,25 \leq \tau_h \leq 36} \hat{\mathbf{D}}_{t,2}^*. \quad (\text{B7})$$

The matrix $\hat{\mathbf{X}}^*$ is then reordered using the Schaake shuffle

$$\mathbf{X}^{*SS_1} = \hat{\mathbf{X}}_{[r]}^* = \left(\hat{\mathbf{X}}_{\bullet,1 \leq \tau_h \leq 24[r_1]}^* \hat{\mathbf{X}}_{\bullet,25 \leq \tau_h \leq 36[r_2]}^* \right). \quad (\text{B8})$$

We now have a matrix \mathbf{X}^{*SS_1} of shuffled and hourly disaggregated forecasts containing an ensemble of $T = 250$ members.

Equations (B1)–(B8) are carried out four times, each time with a different 250 forecasts that are randomly sampled (without replacement) from \mathbf{D}^* .

We now have four calibrated, shuffled, and disaggregated forecast matrices, each containing 250 ensemble members and lead times of 1–36 h. We concatenate these along the ensemble dimension to create a forecast of 1000 ensemble members:

$$\mathbf{X}^{*D} = \begin{pmatrix} \mathbf{X}^{*SS_1} \\ \mathbf{X}^{*SS_2} \\ \mathbf{X}^{*SS_3} \\ \mathbf{X}^{*SS_4} \end{pmatrix}. \quad (\text{B9})$$

For a given cycle, we repeat the entire procedure of Eqs. (B1)–(B9) for all 1157 forecasts in \mathbf{D} , to produce a

set of calibrated, shuffled, and disaggregated forecasts concurrently for all gauges. The process is then repeated for each cycle.

REFERENCES

- Adams, T. E. I., and T. C. Pagano, 2016: *Flood Forecasting: A Global Perspective*. Academic Press, 478 pp.
- Alfieri, L., P. Burek, E. Dutra, B. Krzeminski, D. Muraro, J. Thielen, and F. Pappenberger, 2013: GloFAS – Global ensemble streamflow forecasting and flood early warning. *Hydrol. Earth Syst. Sci.*, **17**, 1161–1175, <https://doi.org/10.5194/hess-17-1161-2013>.
- Andréassian, V., C. Perrin, E. Parent, and A. Bárdossy, 2010: The Court of Miracles of Hydrology: Can failure stories contribute to hydrological science? *Hydrol. Sci. J.*, **55**, 849–856, <https://doi.org/10.1080/02626667.2010.506050>.
- Ballinger, J., B. Jackson, A. Reisinger, and K. Stokes, 2011: *The Potential Effects of Climate Change on Flood Frequency in the Hutt River*. Victoria University of Wellington, 40 pp.
- Bartolini, E., P. Allamano, F. Laio, and P. Claps, 2011: Runoff regime estimation at high-elevation sites: A parsimonious water balance approach. *Hydrol. Earth Syst. Sci.*, **15**, 1661–1673, <https://doi.org/10.5194/hess-15-1661-2011>.
- Beck, H. E., E. F. Wood, M. Pan, C. K. Fisher, D. G. Miralles, A. I. J. M. Dijk, T. R. McVicar, and R. F. Adler, 2019: MSWEP V2 global 3-hourly 0.1° precipitation: Methodology and quantitative assessment. *Bull. Amer. Meteor. Soc.*, **100**, 473–500, <https://doi.org/10.1175/BAMS-D-17-0138.1>.
- Bell, V. A., H. N. Davies, A. L. Kay, A. Brookshaw, and A. A. Scaife, 2017: A national-scale seasonal hydrological forecast system: Development and evaluation over Britain. *Hydrol. Earth Syst. Sci.*, **21**, 4681–4691, <https://doi.org/10.5194/hess-21-4681-2017>.
- Bellier, J., G. Bontron, and I. Zin, 2017: Using meteorological analogues for reordering postprocessed precipitation ensembles in hydrological forecasting. *Water Resour. Res.*, **53**, 10 085–10 107, <https://doi.org/10.1002/2017WR021245>.
- Bennett, J. C., D. E. Robertson, D. L. Shrestha, Q. J. Wang, D. Enever, P. Hapuarachchi, and N. K. Tuteja, 2014: A System for Continuous Hydrological Ensemble Forecasting (SCHEF) to lead times of 9 days. *J. Hydrol.*, **519**, 2832–2846, <https://doi.org/10.1016/j.jhydrol.2014.08.010>.
- Cattoën, C., H. McMillan, and S. Moore, 2016: Coupling a high-resolution weather model with a hydrological model for flood forecasting in New Zealand. *J. Hydrol.*, **55** (1), 1–23.
- , S. Moore, and T. Carey-Smith, 2019: Enhanced probabilistic flood forecasting using optimally designed numerical weather prediction ensembles. Natural Hazards Research Platform

- Contest 2017, 42 pp., <https://www.naturalhazards.org.nz/haz/content/download/14088/74777/file/NHRP%20Contest%202017%20Cattoen%20Final%20Report.pdf>.
- Clark, M., S. Gangopadhyay, L. Hay, B. Rajagopalan, and R. Wilby, 2004: The Schaake shuffle: A method for reconstructing space-time variability in forecasted precipitation and temperature fields. *J. Hydrometeor.*, **5**, 243–262, [https://doi.org/10.1175/1525-7541\(2004\)005<0243:TSSAMF>2.0.CO;2](https://doi.org/10.1175/1525-7541(2004)005<0243:TSSAMF>2.0.CO;2).
- Cloke, H. L., and F. Pappenberger, 2009: Ensemble flood forecasting: A review. *J. Hydrol.*, **375**, 613–626, <https://doi.org/10.1016/j.jhydrol.2009.06.005>.
- Demargne, J., and Coauthors, 2014: The science of NOAA's operational Hydrologic Ensemble Forecast Service. *Bull. Amer. Meteor. Soc.*, **95**, 79–98, <https://doi.org/10.1175/BAMS-D-12-00081.1>.
- Duan, Q. Y., S. Sorooshian, and V. Gupta, 1992: Effective and efficient global optimization for conceptual rainfall-runoff models. *Water Resour. Res.*, **28**, 1015–1031, <https://doi.org/10.1029/91WR02985>.
- Emerton, R. E., and Coauthors, 2016: Continental and global scale flood forecasting systems. *Wiley Interdiscip. Rev.: Water*, **3**, 391–418, <https://doi.org/10.1002/wat2.1137>.
- Gneiting, T., and M. Katzfuss, 2014: Probabilistic forecasting. *Annu. Rev. Stat. Appl.*, **1**, 125–151, <https://doi.org/10.1146/annurev-statistics-062713-085831>.
- , F. Balabdaoui, and A. E. Raftery, 2007: Probabilistic forecasts, calibration and sharpness. *J. Roy. Stat. Soc.*, **69B**, 243–268, <https://doi.org/10.1111/j.1467-9868.2007.00587.x>.
- Gruber, A., and V. Levizzani, 2008: Assessment of global precipitation products. WCRP Series Rep. 128 and WMO/TD-1430, 55 pp., <http://www.wcrp-climate.org/documents/AssessmentGlobalPrecipitationReport.pdf>.
- Hamill, T. M., G. T. Bates, J. S. Whitaker, D. R. Murray, M. Fiorino, T. J. G. Jr., Y. Zhu, and W. Lapenta, 2013: NOAA's second-generation global medium-range ensemble reforecast dataset. *Bull. Amer. Meteor. Soc.*, **94**, 1553–1565, <https://doi.org/10.1175/BAMS-D-12-00014.1>.
- Hamon, W. R., 1973: Computing actual precipitation. Distribution of precipitation in mountainous areas, Vol. 1, WMO Rep. 362, 159–174.
- Jones, D., W. Wang, and R. Fawcett, 2009: High-quality spatial climate data-sets for Australia. *Aust. Meteor. Oceanogr. J.*, **58**, 233–248, <https://doi.org/10.22499/2.5804.003>.
- Li, W., Q. Duan, C. Miao, A. Ye, W. Gong, and Z. Di, 2017: A review on statistical postprocessing methods for hydrometeorological ensemble forecasting. *Wiley Interdiscip. Rev.: Water*, **4**, e1246, <https://doi.org/10.1002/WAT2.1246>.
- Maxey, R., M. Cranston, A. Tavendale, and P. Buchanan, 2012: The Use of deterministic and probabilistic forecasting in Countrywide Flood Guidance in Scotland. *11th BHS National Symp.*, University of Dundee, Dundee, United Kingdom, British Hydrological Society, 7 pp.
- Monhart, S., M. Zappa, C. Spirig, C. Schär, and K. Bogner, 2019: Subseasonal hydrometeorological ensemble predictions in small- and medium-sized mountainous catchments: Benefits of the NWP approach. *Hydrol. Earth Syst. Sci.*, **23**, 493–513, <https://doi.org/10.5194/hess-23-493-2019>.
- Morel, P., 2018: Gramm: Grammar of graphics plotting in Matlab. *J. Open Source Software*, **3**, 568, <https://doi.org/10.21105/joss.00568>.
- Peterson, T., H. Daan, and P. Jones, 1997: Initial selection of a GCOS surface network. *Bull. Amer. Meteorol. Soc.*, **78**, 2145–2152, [https://doi.org/10.1175/1520-0477\(1997\)078<2145:ISOAGS>2.0.CO;2](https://doi.org/10.1175/1520-0477(1997)078<2145:ISOAGS>2.0.CO;2).
- Robertson, D. E., D. L. Shrestha, and Q. J. Wang, 2013: Post-processing rainfall forecasts from numerical weather prediction models for short-term streamflow forecasting. *Hydrol. Earth Syst. Sci.*, **17**, 3587–3603, <https://doi.org/10.5194/hess-17-3587-2013>.
- Rossa, A., K. Liechti, M. Zappa, M. Bruen, U. Germann, G. Haase, C. Keil, and P. Krahe, 2011: The COST 731 Action: A review on uncertainty propagation in advanced hydro-meteorological forecast systems. *Atmos. Res.*, **100**, 150–167, <https://doi.org/10.1016/j.atmosres.2010.11.016>.
- Schefzik, R., 2016: A similarity-based implementation of the Schaake shuffle. *Mon. Wea. Rev.*, **144**, 1909–1921, <https://doi.org/10.1175/MWR-D-15-0227.1>.
- , T. L. Thorarindottir, and T. Gneiting, 2013: Uncertainty quantification in complex simulation models using ensemble copula coupling. *Stat. Sci.*, **28**, 616–640, <https://doi.org/10.1214/13-STS443>.
- Scheuerer, M., T. M. Hamill, B. Whitin, M. He, and A. Henkel, 2017: A method for preferential selection of dates in the Schaake shuffle approach to constructing spatiotemporal forecast fields of temperature and precipitation. *Water Resour. Res.*, **53**, 3029–3046, <https://doi.org/10.1002/2016WR020133>.
- Shrestha, D. L., D. E. Robertson, J. C. Bennett, and Q. J. Wang, 2015: Improving precipitation forecasts by generating ensembles through postprocessing. *Mon. Wea. Rev.*, **143**, 3642–3663, <https://doi.org/10.1175/MWR-D-14-00329.1>.
- Stratton, R. A., and Coauthors, 2018: A Pan-African convection-permitting regional climate simulation with the Met Office Unified Model: CP4-Africa. *J. Climate*, **31**, 3485–3508, <https://doi.org/10.1175/JCLI-D-17-0503.1>.
- Surcel, M., M. Berenguer, and I. Zawadzki, 2010: The diurnal cycle of precipitation from continental radar mosaics and numerical weather prediction models. Part I: Methodology and seasonal comparison. *Mon. Wea. Rev.*, **138**, 3084–3106, <https://doi.org/10.1175/2010MWR3125.1>.
- Tait, A., R. D. Henderson, R. Turner, and X. Zheng, 2006: Thin plate smoothing spline interpolation of daily rainfall for New Zealand using a climatological rainfall surface. *Int. J. Climatol.*, **26**, 2097–2115, <https://doi.org/10.1002/joc.1350>.
- Thielen, J., J. Bartholmes, M. H. Ramos, and A. de Roo, 2009: The European flood alert system – Part 1: Concept and development. *Hydrol. Earth Syst. Sci.*, **13**, 125–140, <https://doi.org/10.5194/hess-13-125-2009>.
- Valéry, A., V. Andréassian, and C. Perrin, 2010: Regionalization of precipitation and air temperature over high-altitude catchments – Learning from outliers. *Hydrol. Sci. J.*, **55**, 928–940, <https://doi.org/10.1080/02626667.2010.504676>.
- Vincent, L. A., and É. Mekis, 2006: Changes in daily and extreme temperature and precipitation indices for Canada over the twentieth century. *Atmos.–Ocean*, **44**, 177–193, <https://doi.org/10.3137/ao.440205>.
- Wang, Q. J., and D. E. Robertson, 2011: Multisite probabilistic forecasting of seasonal flows for streams with zero value occurrences. *Water Resour. Res.*, **47**, W02546, <https://doi.org/10.1029/2010WR009333>.
- , D. L. Shrestha, D. E. Robertson, and P. Pokhrel, 2012: A log-sinh transformation for data normalization and variance stabilization. *Water Resour. Res.*, **48**, W05514, <https://doi.org/10.1029/2011WR010973>.

- , Y. Shao, Y. Song, A. Schepen, D. E. Robertson, D. Ryu, and F. Pappenberger, 2019a: An evaluation of ECMWF SEAS5 seasonal climate forecasts for Australia using a new forecast calibration algorithm. *Environ. Modell. Software*, **122**, 104550, <https://doi.org/10.1016/j.envsoft.2019.104550>.
- , T. Zhao, Q. Yang, and D. Robertson, 2019b: A seasonally coherent calibration (SCC) model for postprocessing numerical weather predictions. *Mon. Wea. Rev.*, **147**, 3633–3647, <https://doi.org/10.1175/MWR-D-19-0108.1>.
- Wellington Regional Council, 1995: Surface water hydrology. Vol. 1, Hydrology of the Hutt Catchment, Wellington Regional Council Rep., 196 pp.
- Woods, R., J. Hendrikx, R. D. Henderson, and A. Tait, 2006: Estimating mean flow of New Zealand rivers. *J. Hydrol.*, **45**, 95–110.
- Zhao, T., J. C. Bennett, Q. J. Wang, A. Schepen, A. W. Wood, D. E. Robertson, and M.-H. Ramos, 2017: How suitable is quantile mapping for postprocessing GCM precipitation forecasts? *J. Climate*, **30**, 3185–3196, <https://doi.org/10.1175/JCLI-D-16-0652.1>.