Coupling forecast calibration and data-driven downscaling for generating reliable, high-resolution, multivariate seasonal climate forecast ensembles at multiple sites

Andrew Schepen*
CSIRO Land and Water, Brisbane, Australia

Yvette Everingham
James Cook University, Townsville, Australia

Quan J. Wang
The University of Melbourne, Melbourne, Australia

Draft manuscript for submission to International Journal of Climatology

*Corresponding Author: Andrew.Schepen@csiro.au

Senior Research Scientist

Phone: +61 7 3833 5513

Post: GPO Box 2583, Brisbane, QLD 4001

Keywords: seasonal forecasting; post-processing; multivariate; forecast verification; bias-correction

This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/joc.6346

This article is protected by copyright. All rights reserved.
Abstract

Calibration and downscaling of ensemble GCM forecasts is becoming increasingly important for hydrological and agricultural modelling in support of the management and protection of valuable natural resources. Moreover, skilful and reliable daily forecast sequences are required to drive decision support models that operate on a daily time step. While downscaling of daily GCM outputs has been developed extensively in climate impacts studies, much less attention has been paid to the downscaling of ensemble GCM forecasts, which has the confronting aspect of low and diminishing skill with increasing lead time. Evidence is building that simple bias correction methods that do not model correlation between forecasts and observations, nor attempt to correct spatial, temporal and inter-variable correlations, produce downscaled forecasts that perform poorly in applications models. Thus downscaling GCM forecasts requires inclusion of a full calibration component to reduce bias, improve reliability, capture skill where it is available and remove negative skill. In this study, we propose a new methodology for coupling a full GCM forecast calibration with empirical methods to (1) instil correct temporal, spatial and inter-variable correlations in ensembles and (2) perform a multivariate downscaling of ensembles to daily sequences at multiple sites. Through a case study application, the proposed methodology is shown to produce skilful monthly–seasonal forecasts of rainfall, temperature and solar radiation at regional spatial scales. It also produces realistic and coherent multivariate daily sequences at multiple sub-grid locations. The new methodology can be applied to more effectively integrate climate forecasts into hydrological and crop models and to support proactive decision making in agriculture and natural resources management.
1. Introduction

Seasonal forecasts have high potential value to agricultural and water resources management industries (Feldman and Ingram 2009; Hansen 2005; Meza et al. 2008). Long-range ensemble climate forecasts extending six or more months ahead are now routinely generated by scientific and governmental agencies using global climate models (GCMs). Publicly available forecasts are normally only issued at coarse spatial and temporal scales, which are sufficient for casual followers of weather and climate. However, to support proactive decision-making in water resources management and agriculture, GCM forecasts need to be translated to the decision maker’s domain through various post-processing techniques, including bias-correction, calibration and/or downscaling (e.g. Manzanas et al. 2018; Maraun 2013; Muchuru et al. 2016; Schepen and Wang 2014; Tian et al. 2014; Wood et al. 2002; Zhao et al. 2017).

Translation to the application domain can involve both spatial and temporal downscaling. For justification, consider crop models, which are usually established for sub-grid (e.g., farm) locations and whose response can be quite sensitive to the distribution of weather within a season (Brown et al. 2018; Hansen 2005). In some applications, downscaled forecasts are required for locations/catchments spanning several grid cells. For example, Potgieter et al. (2005) developed a regional-scale sorghum forecasting system that uses climate analogue inputs at multiple locations. Regional-scale systems can support broad outlooks for industry (e.g. for crop insurance risk), but they do require some care in their development, as Hansen and Jones (2000) point out, because when a crop response is modelled at multiple locations, aggregates can be biased if spatial heterogeneity is not adequately modelled. Consequently, for GCM forecasts to provide inputs to both local and regional-scale systems, it is critical to develop spatial and temporal downscaling methods that contain the meaningful forecast information from the climate model while preserving the statistical properties of the relevant historical daily sequences (Hansen and Jones 2000).

Arguably then, GCM forecasts should be calibrated at an appropriate broad scale of interest before being spatially and temporally downscaled. Calibration of forecasts typically involves using a
statistical model to adjust raw forecasts so that they have little bias, are reliable, and are generally more skilful than climatology (Schepen et al. 2016; Zhao et al. 2017)

Coarse-scale GCM adjustment followed by downscaling is a methodology often used in water resources assessments and forecasting, which may not be surprising, since regional-scale or distributed hydrological modelling requires correctly allocating basin-wide rainfall to sub-catchments to realistically simulate flows. In line with this reasoning, Wood et al. (2002) developed a method to bias-correct coarsely-gridded, monthly climate model outputs of rainfall and temperature, before spatially and temporally downscaling them for use in the VIC semi-distributed hydrological model. The approach has since become widely known as BCSD (bias-correction spatial-disaggregation). Gutmann et al. (2014) reviewed various downscaling methods, albeit for downscaling reanalysis rainfall, and found BCSD still amongst the best performers, especially for the reproduction of extremes and wet day fractions (the proportion of days on which rain is observed). In BCSD, the spatial downscaling is a simple interpolation scheme. The temporal downscaling uses randomly or semi-randomly resampled historical data, e.g., conditioned on wet and dry periods, as a template to derive daily sequences. This data-driven approach to downscaling is appealing, because the daily sequences are consistent with both the large-scale forecasts and the local scale observations.

Several advances upon BCSD have been proposed over the years, albeit mostly in the context of climate impacts studies. Several studies have proposed alternatives to simple interpolation for spatial downscaling. One suggestion is to match the bias-corrected fields to stochastically-generated, spatially-correlated samples in BCSA (bias-correction spatial-analogues) (Hwang and Graham 2013); another is to use multivariate analogues to downscale multiple variables simultaneously from large scale patterns in MACA (multivariate adaptive constructed analogues) (Abatzoglou and Brown 2012).

From a forecasting viewpoint, a major drawback of all the discussed methods (BCSD, SDBC, BCSA and MACA) is that they use quantile-mapping for bias-correction, which is known to be deficient as forecast calibration tool (Brown et al. 2018; Ines and Hansen 2006; Ines et al. 2011; Zhao et al. 2017) and as a downscaling tool (Maraun 2013); largely because it assumes a perfect correspondence between forecasts and observations, regardless of scale. The prevalence of quantile mapping in downscaling stems from the focus on climate change impact studies, where it can be effective as a bias correction tool. Forecasting, however, is quite a different problem with high uncertainty and the added dimension of lead time. Indeed, studies making use of quantile-mapped GCM forecasts in
Australia have found that they perform poorly in agricultural applications (Brown et al. 2018; Western et al. 2018), presumably due to misspecification of temporal, spatial and inter-variable relationships; echoing earlier findings of Ines and Hansen (2006).

As to the complexity of post-processing approaches in active use, simple linear scaling is sometimes seen (e.g. Bazile et al. 2017; Crochemore et al. 2016). Quantile mapping remains popular, despite the aforementioned weaknesses (e.g. Brown et al. 2018; Crochemore et al. 2016; Wetterhall et al. 2015), presumably because it is easy to apply and under the right conditions it can yield substantial improvements over raw forecasts. Other studies have investigated stochastic weather generation. After Ines and Hansen (2006) found that quantile-mapping was unable to correct fundamental problems with their GCM’s rainfall temporal correlation structure, Ines et al. (2011) developed a weather generator to redistribute monthly rainfall across days in a more realistic fashion (quantile mapping was still used initially). For each ensemble member, the weather generator was used to repeatedly generate daily sequences until the monthly rainfall was within 5% of the target, after which a linear adjustment was made to create a perfect match. Such a redistribution resulted in daily rainfall frequencies and intensities that were more realistic for the monthly total rainfall. A limitation of the Ines et al. (2011) approach is that it only sources predictability from GCM rainfall, taking temperature and radiation to be the long term means conditioned on the forecast month and the forecasted occurrence of daily rainfall. For broad applicability, downscaling tools need to be capable of handling a wide range of variables such as rainfall, temperature, solar radiation and evaporation, and manage their associated inter-variable relationships, possibly across multiple sites.

Recent efforts to advance weather generators to downscale forecasts for multiple sites and to use multiple sources of predictability do not seem to have delivered a generally applicable method. Chen et al. (2017) developed a multi-site weather-generator to generate daily precipitation and maximum and minimum temperatures from GCM outputs. While the method could derive multi-site sequences simultaneously for either rainfall or temperature, each variable was still generated independently. More recently, Verdin et al. (2018) developed a parametric stochastic weather generator to output downscaled daily weather sequences based on tercile seasonal climate forecasts from the International Research Institute for Climate and Society (IRI). The approach is flexible enough to incorporate any predictors and, therefore, could theoretically be applied to downscale GCM forecasts. However, the approach is still limited to rainfall as the primary variable, with temperature...
modelled conditionally on the occurrence of rainfall. It appears that stochastic weather-generators require a consolidation of ideas and further development to progress as a general-purpose GCM downscaling tool.

In light of the limitations of existing approaches, we propose to couple a rigorous forecast calibration at coarse-scales with a relatively simple multivariate, data-driven downscaling procedure that simultaneously generates correlated daily sequences for multiple sites. With regards to the calibration of coarse-scale forecasts, we will apply the Bayesian joint probability modelling approach (BJP) (Wang and Robertson 2011; Wang et al. 2009), which has been extensively applied to calibrate GCM outputs of rainfall and temperature (Hawthorne et al. 2013; Schepen et al. 2014; Schepen et al. 2016; Strazzo et al. 2018). Here, we calibrate monthly, spatially-aggregated forecasts. Temporal and inter-variable relationships from historical data will be injected into these monthly forecasts using the Schaake Shuffle (Clark et al. 2004; Vrac and Friederichs 2015). We expect these first steps to work well because BJP-calibrated monthly rainfall forecasts have previously been shown to generate skilful and reliable seasonal streamflow forecasts up to 12 months ahead using a monthly hydrological model (Bennett et al. 2016; Bennett et al. 2017). If BJP generates monthly forecasts with appropriate ensemble spread, then the aim of the downscaling is to preserve the monthly forecast distributions while producing realistic daily sequences with correct weather patterns and extremes. To this end, we combine features of nearest neighbour resampling (Buishand and Brandsma 2001) and the method of fragments (Li et al. 2018; Srikanthan and Mcmahon 2001; Westra et al. 2012) to downscale each ensemble member individually with a suitable historical pattern.

In this study, we present the new methodology, given an acronym for the sake of clarity — FCMD (forecast-calibration multivariate-downscaling) — along with results of an application to ECMWF System4 forecasts in north-eastern Australia. The calibrated coarse-scale forecasts will be checked for skill and reliability. The daily, spatially-downscaled forecasts will be checked for correct temporal, spatial and inter-variable relationships by comparing with observations and raw forecasts. We structure the remainder of the paper as: methods, application and verification, results, discussion and conclusions.
2. Methods

2.1. Forecast Calibration – Multivariate Downscaling

We give a broad overview of the Forecast Calibration – Multivariate Downscaling (FCMD) workflow (Figure 1) here, with detailed information and references given in the subsequent sections 2.2–2.4. The essential FCMD steps are:

1. For a given set of downscaling target locations, e.g., a cluster of weather stations or an array of fine resolution grid cell centroids, GCM forecasts are first interpolated to the set of target locations. Nearest neighbour interpolation is used in this study. Observations and forecasts are then aggregated spatially over all of the locations and in time to the desired resolution. Observations and forecasts are averaged to monthly data in this study. The averaging to larger-spatial scales is intended to reduce noise and to improve the chances of capturing useful GCM climate signals.

2. The aggregated forecasts are calibrated with the BJP modelling approach to reduce bias, produce reliable forecast uncertainty estimates and return forecasts to climatology where skill is worse than climatology. In this study we establish a separate BJP model for each climate variable and lead time (see section 5 for further discussion). Hence if there are \( N \) climate variables and \( L \) lead times to be calibrated, \( N \times L \) BJP models are established. \( M = 200 \) ensemble members are generated by each model. The Schaake Shuffle with a historical data template is then used to connect up the ensemble members across the climate variables and lead times so that the forecast ensembles inherit temporal and inter-variable correlation structures from observations.

3. Each monthly ensemble member is downscaled (disaggregated) to daily sequences and to each target location using a pattern obtained from a historical nearest neighbour (or more accurately, the most similar occurrence). All variables are used in the nearest neighbour search (after standardisation) and the same pattern is applied to all variables. We allow nearest neighbours to be found within a three month window of the target month.
2.2. BJP forecast calibration

Spatially and temporally aggregated forecasts are calibrated using the Bayesian joint probability modelling approach (BJP) (Wang and Robertson 2011; Wang et al. 2009). BJP embeds a model of transformed predictor and predictand variables as a multivariate normal distribution where the transformations allow efficient handling of non-normal variables. It is a general statistical prediction model that can be referred to as a forecast calibration model, a downscaling model or purely as a statistical forecasting model. In this study, BJP predictors are raw monthly GCM forecasts (ensemble means) and observations are monthly observations. Although BJP is described in previous papers, we give a brief description here for completeness.

Denote a generic normalizing transformation function $\psi$ with parameters $\Delta$. The random vector of transformed predictor and predictand variables is $y = (\psi_1(x_1), \psi_2(x_2), \ldots, \psi_N(x_N))$ where $x_i$ is an original untransformed variable. We assume that the joint distribution of a transformed set of variables is multivariate normal, i.e.

$$y \sim N(\mu, \Sigma)$$

where $\mu$ and $\Sigma$ are the means and covariance matrix parameters from the multivariate normal distribution.

For rainfall, $\psi$ is a two-parameter log-sinh transformation (Wang et al. 2012). For other variables, $\psi$ is a single-parameter Yeo-Johnson transformation (Yeo and Johnson 2000). We estimate a single “best” set of transformation parameters for $\psi_i$ using a Bayesian maximum a posteriori (MAP) solution using the method described by Schepen et al. (2016). In contrast to the point estimation of the transformation parameters, the inference of the multivariate normal parameters allows for parameter uncertainty. A Gibbs sampler is used to obtain $M$ samples of $\mu$ and $\Sigma$. In predictive mode, the Gibbs sampler is used to obtain a single sample of $y_{\text{pred}} \mid y_{\text{obs}}, \mu_m, \Sigma_m$ for each of $m = 1, 2, \ldots, M$ where $y_{\text{obs}}$ and $y_{\text{pred}}$ are the random vectors of the BJP predictors and predictands, respectively. The collection of the samples $(y_{\text{pred},1}, \ldots, y_{\text{pred},M})$ constitutes an ensemble forecast.
\( \mathbf{Y}_{\text{ptand}} \) from a BJP model and follows a multivariate normal distribution in transformed space. The forecasts in \( \mathbf{Y}_{\text{ptand}} \) for each transformed predictand variable are back-transformed to the original space using the appropriate inverse transformation \( \psi_i^{-1} \) to obtain \( \mathbf{X}_{\text{ptand}} \), a set of ensemble forecasts in the original (untransformed) space.

Special treatments are involved where a predictor or predictand variable has a physical lower bound, e.g., for rainfall, the lower bound is zero. In the BJP modelling approach, such variables are handled by treating the data as left-censored (Wang and Robertson 2011). The censoring treatment applies to transformation parameter estimation, BJP parameter inference and forecasting.

### 2.3. Schaake Shuffle ensemble reordering

Natural weather patterns indicate that forecast ensemble members traversing time should have a reasonable degree of autocorrelation. For example, it is expected that either warm or cool temperatures can persist for many days and that temperatures exhibit a degree of autocorrelation in general, including at the monthly to seasonal time scales of interest in this study. Similarly, rainfall is expected to have patterns of wet and dry periods. When BJP is applied separately to different variables and months ahead (lead times), the concatenated ensembles will not automatically contain realistic temporal or inter-variable patterns across months by virtue of having little or no correlation between ensemble members with the same indexing position in the ensemble. A further post-processing step is needed to restore realistic covariance in the post-processed ensemble forecasts.

In our work, we choose to use the Schaake Shuffle (Clark et al. 2004), a data-driven procedure, to impose realistic correlation structure in the monthly BJP forecasts. The original implementation of the Schaake Shuffle (Clark et al. 2004) imposed the temporal rank correlation structure of historical observations into forecasts. We describe the essential steps of the procedure here. For a given forecast time period (e.g. month), consider an ensemble forecast of size \( M \) denoted by

\[
\mathbf{X} = (x_1, x_2, \ldots, x_M)
\]

that can be sorted to obtain
\[ \chi = (x_{(1)}, x_{(2)}, \ldots, x_{(M)}) \quad x_{(i)} \leq x_{(2)} \ldots \leq x_{(M)} \]  

Consider also a vector of observations from the historical record for the same time period (e.g. the same month in other years), also of size \( M \)

\[ \mathbf{Y} = (y_1, y_2, \ldots, y_M) \]

that can be sorted to obtain

\[ \gamma = (y_{(1)}, y_{(2)}, \ldots, y_{(M)}) \quad y_{(1)} \leq y_{(2)} \ldots \leq y_{(M)} \]

Furthermore, let \( \text{rank} \) be a function that determines the position of a value from \( \gamma \) in the original unsorted vector \( \mathbf{Y} \). The shuffled forecast ensemble is constructed as

\[ \mathbf{X}_{SS} = (x_{ss,1}, \ldots, x_{ss,M}) \]

where \( x_{ss,q} = x_{(q)} \) and \( q = \text{rank}(\mathbf{Y}, y_{(n)}) \quad n = 1, \ldots, M \). The key to the Schaake Shuffle is the selection of dates used to construct \( \mathbf{Y} \). For the first lead time, the dates may be selected randomly from the historical record. Clark et al. (2004) selected dates from within a 7-day window either side of the target time period. To shuffle subsequent time steps, the dates used to shuffle the first time step are incremented one time step at a time. When the historical observations have high temporal autocorrelation, the result is that the temporal rank correlation structure in labelled ensemble members will be similar to the template data over successive lead times. In this study, to construct \( \mathbf{Y} \), offsets of -30, -15, 0, 15 and 30 days are used to select the first time step of the Schaake Shuffle. Daily data is subsequently aggregated to match the number of days in each monthly forecast to be shuffled.

A neat property of the Schaake Shuffle is that if the same dates are used to construct the template for all forecast variables for a given forecast initialisation date, then both the inter-variable correlations and temporal correlations are reconstructed simultaneously. (The method can also reconstruct spatial correlations in the same way.) Thus the Schaake Shuffle is very useful for producing physically coherent forecasts of multiple variables that are initially calibrated.
independently using a method like BJP. The Schaake Shuffle permits post-processing for high-dimensional problems that would quickly become intractable with fully parametric methods.

Although part of the beauty of the Schaake Shuffle lies in its simplicity, it does have several drawbacks, and, therefore, many modifications have been proposed in recent times. Suggestions include using historical GCM ensembles as the dependence template (Schefzik et al. 2013) or preferentially selecting start dates using analogues or other similarity criteria (Schefzik 2016; Scheuerer et al. 2017; Wu et al. 2018). In our study, we opt to apply the Schaake Shuffle in its original form. One argument for not conditioning the Schaake Shuffle is that the forecast ensembles in seasonal forecasting, in contrast to ensembles in weather forecasting, often represent a wide range of outcomes and therefore it is necessary to include a wide range of observation trajectories in the Schaake Shuffle.

2.4. Nearest neighbour downscaling

Each BJP calibrated forecast contains $M$ ensemble members, each as a sequence of $L$ monthly forecasts. To downscale the forecast to an ensemble of daily sequences at each station, we adapt the method of fragments (MOF), which has mainly been applied for temporal rainfall disaggregation, including from annual-to-monthly to daily-to-sub-daily (e.g. Li et al. 2018; Pui et al. 2012; Srikanthan and Mcmahon 2001; Westra et al. 2012). In some instances it has been applied to disaggregate temperature (Wójcik and Buishand 2003). Our motivation for using MOF is not dissimilar to the motivation for Ines et al. (2011) to use a stochastic weather generator to redistribute GCM rainfall more realistically within a month. That is, it is a way to obtain sequences of daily events that have appropriate magnitudes and frequencies of occurrence for the monthly value.

To our knowledge, MOF hasn’t been previously applied to downscale ensemble climate forecasts and therefore some modifications are required. Here, we apply the method of fragments to downscale all variables simultaneously. Each month is downscaled separately, which raises the possibility of edge effects across month boundaries. We do not expect that the forecast sequences will endure dramatic edge effects since the monthly forecasts have already been shuffled with the Schaake Shuffle. Because it is normally applied as a way to disaggregate observed data (e.g. to infill missing higher-resolution data), MOF usually involves randomly selecting one of $k$-nearest
neighbours according to some probability (e.g. Pui et al. 2012). But because the forecast ensembles generated by BJP are reasonably large ($M = 200$ in this study), representing a continuous density forecast, we take a simpler approach and use the single nearest neighbour as a downscaling template. Choosing the closest nearest neighbour limits any rescaling required to preserve the monthly forecasts. As an (extreme) illustrative example, consider an exceptionally wet monthly rainfall forecast being redistributed with a patently dry sequence; the resulting daily sequence could have unrealistically high rainfall on too few days. Such a problem is noted by Gutmann et al. (2014) in application of the bias-correction statistical-disaggregation (BCSD) approach, requiring limitation of daily values to 150% of previously observed maximums, and a redistribution of remaining rainfall within the month, even when template data is selected conditionally on wet and dry days. We don’t limit the daily rainfall in our approach; rather, we assume that choosing the closest nearest neighbour will keep extreme events in check. However, it remains prudent to check for outliers in the results, which could still be caused by other components in the FCMD workflow (e.g. data transformation).

A nearest neighbour is found by similarity measured using the Euclidean distance for standardised variables. Given $N$ variables of a point forecast $f$ and corresponding observation set $o$, the Euclidean distance between all the points is

$$d_{\text{Euclid}} = \sqrt{\sum_{i=1}^{N} (f_i - o_i)^2}$$

A historical nearest neighbour is found by identifying a historical date for which $d_{\text{Euclid}}$ is the smallest. To avoid the potential problem of variables on different scales skewing the search towards a particular variable when using the Euclidean distance, the variables are standardised to dimensionless quantities as suggested by Buishand and Brandsma (2001). How much of an effect this has will depend on the chosen units of the variables. For example, if monthly rainfall in mm is being compared with monthly average daily minimum and maximum temperatures in degrees C, the benefit of standardisation is likely to be significant because of the possible large variation in rainfall with temperature. However, if the rainfall is a daily rate, then standardisation will have a more subtle effect.
The standardisation here is done on spatially and temporally aggregated data, suggesting that seasonality ought to be considered as a factor. In particular, temperature variations and rainfall frequency and intensity can be vary substantially throughout the year. Therefore, it is deemed unreasonable to search for neighbours from seasons distinct from the forecast period and we only permit neighbours to be found within three months of the forecast target month.

In this study, the standardisation step depends on the type of climate variable. A temperature or solar radiation variable $X_i$ is standardised to $\tilde{X}_i$ by

$$\tilde{X}_i = \frac{X_i - \bar{X}_\text{seas}}{S_{\text{seas}}}$$  \hspace{1cm} (8)

where $\bar{X}_\text{seas}$ is the mean of data available for the search, and $S_{\text{seas}}$ is the corresponding standard deviation. A rainfall variable $X_i$ is standardised to $\tilde{X}_i$ by

$$\tilde{X}_i = \frac{\bar{X}_i}{\bar{X}_\text{seas}}$$  \hspace{1cm} (9)

As mentioned, the method of fragments has mainly been applied to rainfall. Therefore, we detail the procedure for finding a suitable nearest neighbour that can be applied to downscale a set of diverse climate variables simultaneously. Additionally, to our knowledge, the method of fragments hasn’t been previously applied as a spatial downscaling tool, so we also present that straightforward extension here.

For the purposes of the description relevant to this study, it is assumed that variables are already transformed to eliminate mixes of positive and negative values and thus keep the description the same for all variables. For example, temperature data are ensured to be in Kelvin rather than degrees Celsius. With the understanding that the following process is to be repeated with a unique observation set for each ensemble member, fragment weights $\omega$ for a location $s$, day $d$ and variable $i$ are calculated from the observation set as follows:
\[ \omega_{sdi} = \frac{o_{sdi}}{\overline{o}_i} \]  

where \( o_{sdi} \) are the individual daily observed values and \( \overline{o}_i \) is the average spatially over \( K \) locations in the region and temporally over \( T_{agg} \) preceding days.

Once weights have been calculated, a new forecast for the spatially and temporally averaged target, \( \overline{f}_i \), is downscaled to each location and to daily values as follows:

\[ f_{sdi} = \omega_{sdi} \overline{f}_i \]  

The downscaled forecast is designed to take on the behaviour of observations because the historical data template preserves temporal, spatial and inter-variable correlation structures (within the aggregation period of length \( T_{agg} \)). As with all data-driven methods, a short data record is potentially problematic when searching for nearest-neighbours. Especially with a seasonality restriction, since the same observed daily sequences could be selected repeatedly. To increase the diversity in the selected sequences, given that the sequences are only to be used as a template for downscaling, rolling aggregates of the observed data are allowed in the search, where the rolling aggregates are prepared for periods of 28–31 days, the possible number of days in a month. This not only ensures a large number of available samples, but also introduces timing shifts into the forecast ensembles. In the undesirable case that similar forecast ensemble members are downscaled using the same sequence, they will still differ by proportionality. Occasionally, enforcements are needed so that sensible sequences are chosen. For example, if forecast rainfall is a positive amount then it is not reasonable to choose a neighbour that has zero rainfall.

3. Application and verification

3.1. Study area and data

Six study weather stations are chosen in the Burdekin region of Australia (Figure 2), which is currently an active area of agricultural research. For these sites, we choose to focus on long-range forecasts issued in September, running 6 months to February, which is typically a sugar growing...
period, although in this study we develop and examine efficacy of the FCMD method and reserve application results for other studies.

Daily observations of rainfall (Precip; mm), minimum temperature (Tmin; °C), maximum temperature (Tmax; °C) and radiation (Srad; MJ/m²/day) are obtained from the Silo dataset for each of the selected stations. The dataset and variables are commonly used by crop modellers in Australia. The data have a discretised nature with Precip reported to the nearest 0.1 mm, Tmin and Tmax to the nearest 0.5 °C and Srad to the nearest 1.0 MJ/m²/day. We note that solar radiation data might not be available in other applications of FCMD; in such cases Antonanzas-Torres et al. (2013) review a range of simple and complex methods for estimation of solar radiation that could be used, including some methods that require only temperature and/or rainfall.

GCM forecasts are obtained from ECMWF’s System4 (Sys4; Molteni et al. 2011) seasonal forecast system at a daily time step and on a 0.75 degree regular grid for the period 1981–2016. To prepare the data for calibration, GCM forecasts are spatially aggregated by interpolating the GCM forecasts to each point, by nearest neighbour, and then averaging across the six stations. In other words, the aggregated GCM forecast is weighted by the number of stations within the grid cell. The corresponding Silo observations are simply averaged across the six stations.

3.2. Forecast verification

3.2.1. Cross-validation

Leave-one-year-out cross validation is applied to generate and evaluate the study results. That is, for each forecast month, data from the verifying month is omitted from the model fitting part of BJP calibration, plus data in the relevant window is omitted from the Schaake Shuffle and nearest-neighbour downscaling procedures. While leave-one-year-out cross-validation is standard practice for seasonal climate forecast evaluation where hindcast datasets are relatively short, if multi-year auto-correlation exists it would be prudent to eliminate additional years to avoid inflating estimates of forecast skill. One-year-lag autocorrelation was calculated for each variable and target month with no evidence of significant autocorrelation found, indicating leave-one-year-out is sufficient for our case study.
3.2.2. Seasonal skill and reliability evaluation

Reliability is the property of statistical consistency between forecasts and observations. A reliable forecasting system will accurately estimate the likelihood of an event. Reliability is checked by analysing the distribution of probability integral transformations or PIT values (Gneiting et al. 2007).

The PIT for a forecast CDF \( F_t \) for event \( t \) and paired observation \( o_t \) is defined by

\[
\pi_t = F_t(o_t)
\]  

(12)

If a forecasting system is reliable and the forecasts are continuous, then the PIT values for a set of forecasts will follow a standard uniform distribution. Hence, we check for uniformity visually using the PIT uniform probability plot, otherwise known as predictive Q-Q plot, by plotting sorted PIT values \( \pi_{(i)} \) against theoretical uniform quantiles \( u_i \) for \( i = 1, \ldots, T \) (Renard et al. 2010; Wang and Robertson 2011). As a more formal test of uniformity we plot Kolmogorov-Smirnov confidence bands on the PIT uniform probability plot (Laio and Tamea 2007). The test is passed if no PIT values lie beyond \([u_i - c(\alpha)\sqrt{2/T}, u_i + c(\alpha)\sqrt{2/T}]\) where \( c(\alpha) = 1.224 \) for a 95% confidence test.

Forecast skill is evaluated by comparing the continuous ranked probability score (CRPS; Matheson and Winkler 1976) for different model forecasts. The CRPS for a given forecast and observation is defined as

\[
\text{CRPS}_t = \int \left[ F_t(y) - H(y - o_t) \right]^2 dy
\]  

(13)

with \( H(y - o_t) = \begin{cases} 0 & \text{if } y < o_t \\ 1 & \text{if } y \geq o_t \end{cases} \)

where \( F_t \) is the forecast CDF for event \( t \); \( o_t \) is the observed value; and \( H \) is the Heaviside step function. The CRPS puts weight on both forecast accuracy and reliability in ensemble spread.

The average CRPS for two sets of forecasts is compared to calculate calculating the CRPS skill score.
CRPS_{SS} = \frac{\overline{CRPS}_{\text{ref}} - \overline{CRPS}}{\overline{CRPS}_{\text{ref}}} \times 100 \quad (\%) \quad (14)

where the overbar indicates averaging across a set of events, \overline{CRPS} is the average CRPS for the focus forecasts and \overline{CRPS}_{\text{ref}} is the average CRPS for a set of reference forecasts. In this study, reference forecasts with ensemble size matching the BJP forecasts are obtained by sampling from marginal distributions fitted to monthly data. To obtain valid references for seasonal forecast verification, the monthly reference forecasts are also Schaake Shuffled.

3.2.3. Validation of downscaled daily sequences

Daily forecasts are validated in terms of having realistic distributions and covariance structure. For each location and variable, the magnitudes of the daily forecast values are visually checked by comparing the [0.25, 0.75] and [0.05, 0.95] quantile ranges of daily forecast values and Silo observations. For rainfall, the frequency of wet days is checked by comparing the proportion of wet days (wet%) where Precip exceeds 1mm. For the examination of quantile ranges and wet day fractions, an average is taken over the \( M \) parameter sets so that the same number of data points are used in the comparisons.

Kendall correlations, as a measure of rank correspondence, are calculated to evaluate temporal, spatial, and inter-variable relationships. Comparisons are made between the downscaled forecasts, raw Sys4 forecasts and Silo observations to determine any discrepancies in the correlation structures of raw and downscaled forecasts.

For temporal correlations, lags of 1–7 days are evaluated for each station and variable. For spatial correlations and inter-variable correlations, the correlation is calculated on all possible pairwise combinations. Correlations are averaged over \( T \) years. For forecasts, a further average is taken over the \( M \) ensemble members.
4. Results

4.1. Skill and reliability of BJP-calibrated forecasts

CRPS skill scores for monthly and seasonal forecasts are shown in Figure 3. The seasonal forecasts are obtained from the monthly forecasts by aggregating each ensemble member over three months. All forecasts are issued on the 1st of September. For monthly forecasts, the target months Sep–Feb correspond to lead times of 0–5 months. Similarly, for the seasonal forecasts, the target seasons SON–DJF correspond to forecasts with 0–3 months lead time. The skill scores measure relative reduction of CRPS error relative to the reference forecast (see section 3.2.2) over the period 1981–2016. The CRPS skill scores shown are estimates based on only 36 events. Bootstrap resampling estimates indicate that the CRPS skill scores have significant uncertainty. The standard deviation of bootstrapped CRPS skill scores ranges from approximately 1-2 percentage points for variables and months up to 9 percentage points where skill is strongly positive.

Monthly forecasts of Tmin, Tmax and Srad are skilful for the first month, and more skilful than the corresponding Precip forecasts. Beyond the first month, longer lead-time skill for is evident for Nov forecasts. Longer lead time skill for Nov rainfall forecasts has been previously observed in the study region using other GCMs and observational datasets (Hawthorne et al. 2013). While no study appears to have confirmed it, the skill is possibly related to the timing of the onset of the wet season. Seasonal forecasts of all variables are more skilful than monthly forecasts. Viewing the monthly forecasts as two distinct periods, SON and DJF, illustrates that the first 3 months of the forecast are moderately skilful whereas only a small amount of skill is available in the second 3 months of the forecast. BJP calibration is designed to push forecasts towards climatology where skill is low. However, some small negative skill scores are observed, particularly in monthly forecasts, where there is essentially no skill and because calibration is undertaken in a cross-validation mode.

PIT uniform probability plots for the monthly and seasonal forecasts are shown in Figure 4 to assess reliability. All of the lead times for a given variable are plotted on the same graph using different colours. It is evident that both the monthly and the seasonal forecasts are overall reliable in terms of ensemble spread. None of the PIT values fall outside the Kolmogorov-Smirnov 95% confidence
bands, indicating that the test for uniformity passes, thus providing more formal evidence that the forecasts are statistically reliable.

4.2. Distributions of daily values

The distributions of daily calibrated forecast values are compared with Silo observations for the drier months, Sep–Nov (Figure 5), and the wetter months, Dec–Feb (Figure 6). Stations are presented in columns and variables are presented in rows. For each variable, the all-station median is plotted as a grey dashed line to help discern shifts in the central tendency of the distributions between stations. For rainfall, the median is determined for wet days only, and the proportion of wet days (wet%) is calculated separately.

Overall the distributions of the forecast and observed values have approximately equal quantiles across the range of cumulative probabilities as well as the correct distributional shape. For example, Tmin and Tmax are approximately normally distributed, Precip is positively skewed whereas Srad is negatively skewed. The forecasts and observations appear to follow the same trends between locations. For example, for Precip during Dec–Feb, station 33090 has the highest proportion of wet days yet has the lowest daily rainfall amounts. Distributions of daily rainfall match closely for forecasts and observations, demonstrating that the downscaling method is reliable for generating realistic rainfall values with a realistic frequency of occurrence and intensity on any given day.

The distributions of daily values appear to accurately reflect the station location in some instances, in both forecasts and observations. For example, compare Tmin and Tmax for the inland station 33090 to the coastal station 33035. The median Tmax at the inland station 33090 is > 1 degree above the all-station median during SON whereas the median Tmax forecast for the station 33035 is below the all-station median. The differences in maximum temperatures can be justified in the context of sea-breezes limiting the maximum daytime temperatures at the coastal location. In contrast, the Tmin values at station 33090 indicate the coldest overnight temperatures are observed there while warmer overnight temperatures are observed at 33035.
4.3. Temporal correlations

We compare the average temporal correlations in daily FCMD forecasts with observations and with raw Sys4 forecasts (Figures 7-8). Average Kendall correlations for lags of 1–7 days are shown for each variable and location. Figure 7 is for Sep–Nov and Figure 8 is for Dec–Feb. In all cases the correlation tends to reduce as the number of days lag increases, however the rate of decline varies. It is noted that because some of the locations reside within the same grid cell, the correlations in raw GCM forecasts will the same for those locations (the spatial correlation is 1 – see next section).

Referring to Silo observations, the temporal correlation behaviour varies substantially between variables and time of year. Temporal correlation in Tmin and Tmax is quite strong and persistent during Sep-Nov, reducing from 0.6 to 0.4 as the lag increases from 1–7 days. However, during Dec-Feb, temporal correlations in Tmax and Tmin are generally weaker and less persistent. Similar comments apply to Srad. In contrast, temporal correlations in Precip show some opposite behaviour. 1-day lag correlations in Precip are slightly lower in Sep-Nov (0.2-3) compared to Dec-Feb. Precip temporal correlation tends to 0 at about 3 days lag during Sep-Nov and at about 4-5 days lag during Dec-Feb. In summary, Precip is seen to be more persistent during the wet period and temperature is seen to be more persistent during the drier period.

It is immediately clear from Figure 7 (Sep-Nov forecasts) that while temporal correlations in raw GCM forecasts have a similar broad behaviour to Silo observations, the magnitudes of the correlations in raw forecasts tend to be higher, a result that is consistent with previous studies. There are exceptions, however. For Tmin, the raw forecasts have stronger temporal correlations than Silo at some locations and weaker temporal correlations at other locations. Figure 8 (Dec-Feb forecasts) shows a much closer match between raw GCM and Silo observations for correlations in Tmin, Tmax and Srad. However, there is a tendency for the correlations to be higher at longer lags, and also the correlation in Precip is still too strong.

The FCMD forecasts exhibit temporal correlations that are much more consistent with Silo observations overall. For all variables and locations the temporal correlations in FCMD forecasts are much improved compared to raw GCM forecasts.
4.4. Spatial correlations

Spatial correlations are compared between FCMD forecasts, Silo observations and raw Sys4 forecasts (Figure 9). Average Kendall correlations are plotted for each variable and for each pairwise combination of stations (within each plot). We do not necessarily expect that the coarse spatial GCM grid correctly captures the spatial correlation between stations. Indeed, for stations within the same grid cell, all raw Sys4 forecasts will be perfectly correlated when a nearest neighbour interpolation is used to obtain the station values. For Tmin, Tmax and Srad the spatial correlations in raw Sys4 forecasts can be stronger or weaker than in observations. In contrast, for Precip, the spatial correlation is always stronger relative to observations. FCMD downscaling reproduces the observed spatial correlations much more closely. The results are not perfect for Tmin, Tmax and Srad with FCMD spatial correlations being slightly lower than observed.

4.5. Inter-variable correlations

Inter-variable correlations are compared between FCMD forecasts, Silo observations and raw Sys4 forecasts (Figure 10). Average Kendall correlations are plotted for each pairwise combination of variables and for each location (within each plot). As with temporal correlations, the raw Sys4 correlations are identical for locations that are in the same grid cell.

In some instances, the raw Sys4 forecasts have approximately correct inter-variable correlations. For example, the raw Sys4 forecasts capture the little or no correlation between Tmax and Precip and, for some locations, the low correlation between Tmin and Srad. However, in other cases, the raw GCM inter-variable correlation is too strong. For example, at some locations, the correlation between Tmin and Tmax is above 0.6 when the correlation in observations is closer to 0.5. The correlation between Precip-Srad appears to be too strong in general for all locations in the raw Sys4 forecasts. In contrast, the FCMD inter-variable correlations are similar to Silo observations for all pairwise combinations of variables and locations.
5. Discussion

In this study, we compare the correlation structures in raw forecasts with FCMD forecasts. Figures 7–8 largely support previous conclusions that raw GCM forecasts have autocorrelation structures that are too strong when compared to observations (Brown et al. 2018; Ines and Hansen 2006; Ines et al. 2011). However, we also find that the raw Sys4 forecasts sometimes have autocorrelation that is similar to, or weaker than, observations. Additionally, we show that spatial and inter-variable correlations are often wrongly specified in raw GCM forecasts (Figures 9–10). Our proposed FCMD method is able to generate downscaled (spatially and temporally) forecast ensemble members that have the correct temporal, spatial and inter-variable correlation structures. It would be very valuable to determine whether the improved characteristics of FCMD forecasts translate into improved agricultural forecasts and ameliorates problems identified in previous studies that use simpler methods that inherit GCM correlations structures (e.g. quantile mapping). It is possible that correcting correlations in climate ensembles will still not be enough, and that further post-processing of streamflow or yield predictions will be required.

As described in section 2.4, no constraint is placed on the values that can be generated in the nearest neighbour downscaling, thus it is possible that unreasonable values could be generated by the chaining of statistical functions. To check the reasonableness of the downscaled values, particularly extremes, we plotted the value of each downscaled forecast against its nearest neighbour used in the downscaling (not shown). While scaling can occur, it appears that appropriate weather patterns are overwhelmingly selected for the downscaling and therefore scaling is limited. Extrapolation occurs within reasonable limits, confirming that filtering of extreme values is not urgently needed, at least not in the current application. Related to this, future studies may test a more relaxed selection of nearest neighbours (e.g. sample from K-nearest neighbours).

We identify that there may be ways to improve upon the fundamental construction of FCMD. For example, while effective, it is not necessarily optimal to calibrate monthly forecasts independently for each variable with BJP, then inject the multivariate relationships (with the Schaake shuffle) before downscaling. We hypothesise that skill and spread could become quite different through a joint calibration of all variables. By way of example, consider the monthly Srad forecasts, which have low or slightly negative skill during Dec–Feb (Figure 3). Within the BJP calibration step, these
forecasts are returned to an unconditional climatology, whereas it might be more reasonable to return a climatology conditional on the forecasts of the other variables where there is inter-variable correlation (Figure 10) and skilful forecasts of the related variables (Figure 3). Further investigation is needed to uncover whether such a correction is warranted. For now, we note that Piani and Haerter (2012) found that jointly bias-correcting climate model rainfall and temperature in a 2D copula framework afforded some benefits over treating each variable separately. Theoretically, BJP is able to calibrate many variables jointly, and so the extension will be straightforward, however such an approach would need to be rigorously evaluated for its merits in forecasting, given that increasing the number of parameters in the calibration model is likely to increase the risk of overfitting.

The provision of domain-specific forecasts opens up opportunities to tailor forecasts to specific applications. The BJP calibration component could also be augmented to predict variables that aren’t output by the GCM, e.g. potential evaporation (Zhao et al. 2019). Future work should investigate additional ways to improve skill and provide additional output variables. For example, the BJP calibration component of FCMD can be augmented to include other climate predictors from the GCM, related to climate drivers such as an ENSO (e.g. Schepen et al. 2014; Schepen et al. 2016).

FCMD is designed for seasonal forecasting and therefore it is not expected that forecasts from the first weeks would be suitable to inclusion in a hydrological or crop model for short term forecasting. Alternative approaches focusing on calibration of daily or multi-week forecasts directly (Schepen et al. 2018) are arguably needed, especially where forecasts may be highly skilful initially and tied to the GCM initial conditions.

Keeping on with alternative approaches, spatiotemporal relationships may be introduced into the forecasts by explicit/parametric modelling. Such extensions can quickly become intractable due to increases in the number of parameters and model complexity, however, approaches including Bayesian hierarchical models are worthy of investigation, particularly in conjunction with dimensionality-reducing methods. Our intuition is that the while development of parameterised approaches is a long-term goal, hybrid post-processing systems involving pragmatic approaches like the data-driven downscaling method in FCMD, will continue to have a role in linking GCM forecasts and applications in the near term.
The water resources management and agricultural sectors often want to incorporate GCM forecasts into their modelling tools and decision support systems. Seasonal forecasting centres tend to issue coarsely-gridded outlooks with a supporting narrative, which are informative for casual forecast users. However, calibrated and downscaled daily forecast sequences are required for quantitative modelling using hydrological models and crop models.

A significant amount of literature has developed methods for bias-correcting and spatially or temporally downscaling GCM outputs or reanalysis data, mainly for climate impacts studies. In climate impacts studies there is no strict lead time dimensionality and simulations are not always synchronous with observations. In forecasting, skill is highly variable with forecast lead time, as well as with time of year and location. Accordingly, forecasting involves ensembles to represent a range of possible forecast trajectories. Even at coarse model resolutions, it is essential that forecasts are properly calibrated to reduce bias, improve ensemble spread, capture model skill and filter out negative skill, all while improving the representation of temporal, spatial and inter-variable correlations. Downscaling the forecast information to the local domain is a significant further challenge since the calibration of the coarse-scale forecast must be preserved.

Researchers and practitioners alike have recognised partial-calibration methods such as quantile-mapping are deficient for GCM forecast calibration and downscaling, especially for applications in hydrological and crop modelling. In our study, we developed a new methodology, termed FCMD (forecast calibration – multivariate downscaling) that sought inspiration from current approaches like BCSD (bias correction – spatial disaggregation), and built a workflow to couple full forecast calibration with empirical downscaling approaches. We introduce a modified nearest-neighbour downscaling approach to downscale individual ensemble members temporally (from monthly to daily) and spatially (from regions to points).

FCMD captures GCM skill at aggregated spatial and temporal scales where climate signals tend to be stronger. In this study, the focus was on the direct calibration of rainfall, temperature and solar radiation using BJP, however FCMD is easily extensible to incorporate prediction of other variables.
of interest as well as other sources of predictability. FCMD forecasts were calibrated and monthly
time scales and were shown to be reliable at both monthly and seasonal time scales.

FCMD downscaled forecasts were demonstrated to have approximately correct temporal, spatial
and inter-variable correlations. Indeed, the correlation structures of FCMD forecasts are much
improved compared to raw GCM forecasts and, by extension, forecasts corrected with techniques
like mean bias correction and quantile-mapping. The development of calibrated and downscaled
forecasts with realistic correlation structures at both monthly and daily time scales through FCMD
has the potential to significantly improve the uptake of GCM forecasts in agriculture and water
resources management. Work is underway to rigorously evaluate FCMD forecasts for long-range
hydrological forecasting and hydrological modelling.
7. References


This article is protected by copyright. All rights reserved.


Manzanas, R., and Coauthors, 2018: Dynamical and statistical downscaling of seasonal temperature forecasts in Europe: Added value for user applications. *Climate Services*, 9, 44-56.


**Figure Captions**

**Figure 1**: Forecast-calibration multivariate-downscaling (FCMD) work flow to produce daily, downscaled ensemble forecast sequences from monthly GCM forecasts.

**Figure 2**: The location of the six study weather stations in north-eastern Australia (blue dots) and the ECMWF System4 grid cell boundaries (grey dashed lines).

**Figure 3**: CRPS skill scores for BJP-calibrated monthly and seasonal forecasts issued in September. Skill scores are calculated against climatological reference forecasts for the years 1981–2016.

**Figure 4**: PIT reliability plots for BJP-calibrated monthly and seasonal forecasts issued in September. Different coloured points represent different lead times from 0–5 months for monthly forecasts and
0–3 months for seasonal forecasts. The dashed-grey lines are the Kolmogorov-Smirnov significance bands.

Figure 5: Boxplots comparing the distribution of daily forecast values with the distribution of daily observed values for observations (O) and forecasts (F) during Sep-Nov. Boxplots show the IQR and the \([0.05,0.95]\) quantile ranges. For rainfall, the distribution is plotted for wet days only (\(\text{Precip} \geq 1\text{mm}\)) and the proportion of wet days is also given (\text{wet}%).

Figure 6: As for Figure 5, except for Dec-Feb forecasts and observations.

Figure 7: Average temporal correlations in Silo observations, FCMD forecasts (cal) and raw Sys4 forecasts for the Sep-Nov months. Kendall correlations are calculated for lags of 1–7 days.

Figure 8: As for Figure 7, except for the Dec-Feb months.

Figure 9: Comparison of average spatial Kendall correlations in Silo observations, FCMD forecasts and raw Sys4 forecasts for the Sep-Nov months. Each marker represents a station pairing (15 combinations).

Figure 10: Comparison of average inter-variable Kendall correlations in Silo observations, FCMD forecasts and raw Sys4 forecasts for the Sep-Nov months. Each marker represents a single location.
Distribution of daily values during SON

JOC_6346_Figure5.jpg
Distribution of daily values during DJF

33035

33069

33002

33033

33122

33090

Tmin (°C)

Tmax (°C)

Precip (mm.day⁻¹)

Sed. (Mg.m⁻².day⁻¹)

wet%

O: 27.7; F: 27.1

O: 30.1; F: 29.2

O: 27.1; F: 26.7

O: 25.2; F: 25.6

O: 24.6; F: 24.6

O: 32.6; F: 31.8

JOC_6346_Figure6.jpg
This article is protected by copyright. All rights reserved.
A novel methodology (see figure) is developed to couple a rigorous GCM forecast calibration with data-driven downscaling to generate high-resolution and multivariate climate forecasts in the form of ensembles.

Skilful and reliable seasonal forecasts are downscaled using historical data patterns to obtain much more realistic temporal, spatial and inter-variable relationships compared to raw forecasts.

The new forecasts are promising to enable better use of GCM forecasts in agriculture and water resources management where existing methods result in unreliable linkages and poor forecasts.
Author/s:
Schepen, A; Everingham, Y; Wang, QJ

Title:
Coupling forecast calibration and data-driven downscaling for generating reliable, high-resolution, multivariate seasonal climate forecast ensembles at multiple sites

Date:
2020-03-30

Citation:

Persistent Link:
http://hdl.handle.net/11343/286480

File Description:
Accepted version